

SPACE TECHNOLOGY SUMMER INSTITUTE

June 17 - July 29, 1966

Presented by

THE UNIVERSITY OF SOUTHERN CALIFORNIA

University College, Noncredit Programs Office

and

School of Engineering, Department of Aerospace Engineering

Sponsored by

THE NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

TABLE OF CONTENTS

ACKNOWLEDGEMENT	i
FOREWORD	ii
STAFF	iii
STUDENTS	iv

TOPICS

SECTION

"SPACE APPLIED MECHANICS"	I ✓
R. H. Edwards	
"SPACE SYSTEMS ENGINEERING"	II ✓
Kane Cassani	
"SPACE SCIENCE"	III ✓
Roman K. C. Johns	
"SPACE GUIDANCE AND CONTROL"	IV ✓
Nasser Nahi	
"SPACE COMMUNICATIONS"	V ✓
R. Scholtz and C. Weber	

ACKNOWLEDGMENT

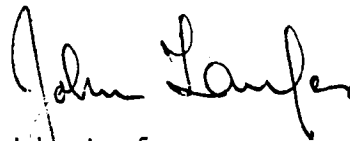
This Institute is sponsored by the National Aeronautics
and Space Administration, under contract # NSR 05-018-055.

FOREWARD

Because of its demand for both scientific depth and broad inter-disciplinary perspective, the field of Space Technology offers a unique opportunity and challenge to the scientist-engineer.

We have seen remarkable gains in the last ten years in such diverse areas as scientific program management or Systems Analysis, and in the development of small highly reliable electronic equipment such as that used in communication satellites. In both examples, the impetus for the gains came largely from requirements for space systems, but the resulting benefits will be felt in many unrelated fields.

Today, Space Technology represents a focal point in the development of highly sophisticated scientific systems, and we hope that this Summer Institute will be of value in delineating the problem areas and in indicating some of the exciting prospects of a future in the field.

A handwritten signature in black ink, appearing to read "John Laufer", with a long, sweeping flourish extending from the end of the name towards the right margin.

John Laufer
Professor and Chairman
Department of Aerospace Engineering

Space Technology Summer Institute

STAFF

DIRECTOR

Edwards, R. H. Professor of Aerospace Engineering
University of Southern California

LECTURERS

Adams, James Supervisor High Impact and Advanced
Mechanisms Group
Jet Propulsion Laboratory

Casani, Kane Engineering Group Supervisor
Jet Propulsion Laboratory

Johns, Roman Professor of Physics
Loyola University

Nahi, Nasser Associate Professor of Electrical
Engineering
University of Southern California

Scholtz, Robert Assistant Professor Electrical
Engineering
University of Southern California

Weber, Charles Research Associate, Electrical
Engineering
University of Southern California

SPECIALIST LECTURER

Kyner, Walter T. Professor of Mathematics
University of Southern California

TECHNICAL CONSULTANT

Laufer, John Professor of Aerospace Engineering
University of Southern California

STUDENTS

Aley, Richard P. California State Polytechnic College
San Luis Obispo, California

Rt. 2, Box 425
San Luis Obispo, California 93401

Alward, Joseph Francis Jr. Sacramento State College
Sacramento, California

1719 Pluto Way
Sacramento, California

Bair, Raymond Everette Washington State University
Pullman, Washington

North 4003 Hawthorne Street
Spokane, Washington 99208

Bird, Richard E. Utah State University
Logan, Utah

303 South 1st West
Logan, Utah 84321

Cunningham, Jerome Earl Fresno State College
Fresno, California

2011 N. Thorne
Fresno, California 93705

Dalich, Stephen John Fresno State College
Fresno, California

2027 N. Orchard #F
Fresno, California

Dubisch, Russell John Reed College
Portland, Oregon

19248 93rd Place W.
Edmonds, Washington 98020

duVigneaud, Jean Louis..... University of Santa Clara
Santa Clara, California

6621 Del Cerro Blvd.
San Diego, California 92120

Elliott, John Douglas Portland State College
Portland, Oregon

18408 S.E. Ivon Court
Gresham, Oregon

STUDENTS (cont.)

Gorger, Donald Edward University of Portland
Portland, Oregon

6941 North Haven Street
Portland, Oregon 97203

Graddy, Joseph Craig University of California at Los Angeles
Los Angeles, California

1526 Sorrento Dr.
Pacific Palisades, California 90272

Griffith, Jerry Lee Alaska Methodist University
Anchorage, Alaska

Box 8-635
Mt. View, Alaska 99504

Groleau, Michael Keith College of Guam
Guam

1212 A Mango Drive
APO San Francisco, California 96334

Haynes, James Ray California State Polytechnic College
San Luis Obispo, California

P.O. Box 124
Oceano, California 93445

Heiser, John Emerson Harvey Mudd College
San Bernardino Valley, California

2905 Lugo Avenue
San Bernardino, California 92404

Holberg, Jay Brian Whitworth College
Spokane, Washington

10220 North College Road
Spokane, Washington 99218

Howard, Richard Alan California State College at Long Beach
Long Beach, California

621 So. Pannes Ave.
Compton, California 90221

Hubner, Mike Frank University of Southern California
Los Angeles, California

216 North Ynez Street
Monterey Park, California 91754

STUDENTS (cont.)

Hughey, Lee Van College of Guam
Agana, Guam

Trust Territory Hq.
Saipan, Mariana Is. 96950

Inouye, Lance Masao Stanford University
Stanford, California

3554 Kumu Street
Honolulu, Hawaii 96822

Jewett, Robert Allen University of Puget Sound
Tacoma, Washington

8649 A Street
Tacoma, Washington 98444

Knopf, Robert G. University of Southern California
Los Angeles, California

5308 Weatherford Drive
Los Angeles, California 90008

Kozman, Theodore Albert University of Southern California
Los Angeles, California

2301 Cartlen Drive
Placentia, California

Lantz, Paul Robert Seattle University
Seattle, Washington

4536 S.W. Director
Seattle, Washington 98116

Noxon, Arthur M. Fresno State College
Fresno, California

1143 So. Minnewawa
Fresno, California

Okazaki, Alan Shizvo California State Polytechnic College
San Luis Obispo, California

20 South Main Street
Lodi, California 95240

Olson, Wayne Marvin University of Washington
Seattle, Washington

Rt. 2 Box 279
Mt. Vernon, Washington 98273

Pereira, David Benjamin	San Francisco State College San Francisco, California 215 Arroyo Drive Pacifica, California 94044
Sergeant, Albert James III	Alaska Methodist University Anchorage, Alaska 30-332-D Cherry Drive Elmendorf A.F.B., Alaska 99504
Spanos, William J.	La Sierra College La Sierra, California 1154 W. Highland Ave. Redlands, California
Sparks, Ernest Leroy	Portland State College Portland, Oregon 1525 S.W. 10th Avenue #5 Portland, Oregon
Toda, Alvin Earl	The University of Hawaii Honolulu, Hawaii 1363 Hoowali Street Pearl City, Hawaii 96782
Theriault, Wesley R.	University of Southern California Los Angeles, California 9411 C San Juan South Gate, California
Vindum, Jorgen Ole	University of California at Berkeley Berkeley, California 1181 Alicante Drive Pacifica, California 94044
Wright, Robert Keith Jr.	La Sierra College Riverside, California 742 Alta Vista Drive Vista, California 92083

N67. 8.046.2

SPACE SCIENCE - "SPACE APPLIED MECHANICS"

I

by

R. H. Edwards

PREFACE

This is one of six courses given in the Third Space Technology Summer Institute. The purpose of this course is to present an introduction to the topics of Space Trajectories and Propulsion. In previous years, a separate course in propulsion was given, but the material was combined with orbital mechanics this year in order to allow for courses in both Spacecraft Mechanical Design and Space Systems Engineering: The primary purpose of the course was to give sufficient background in the two areas so that the design parameters could be understood, but detailed descriptive material was not presented.

Qualitative material was available through the use of reading assignments from Glasstone,¹ which was furnished to the students.

The basic material presented in these notes is available in many textbooks in orbital mechanics and gas dynamics. The notation used here is largely similar to that of Berman,² while more detail in the gas dynamics area can be found in such books as Liepman and Rashko.³ The discussion of optimum launch procedures used can be amplified greatly by reading the Section by Freed in Space Technology.⁴ The material in Gemini Rendezvous is from a N.A.S.A. paper by Culpepper and Bordano.⁵

REFERENCES

1. "Sourcebook on The Space Sciences," Samuel Glasstone
D. Van Nostrand Co. Inc. Princeton, New Jersey. 1965
2. "The Physical Principles of Astronautics," Arthur Berman
John Wiley and Sons Inc. New York. 1961
3. "Elements of Gas Dynamics," H.W. Liepman and A. Rashko
John Wiley and Sons Inc. New York. 1957
4. "Space Technology," edited by Howard S. Seifert
John Wiley and Sons Inc. New York. 1959
5. "Orbital and Rendezvous Report for Gemini IX," Bobby K. Culpepper
and Aldo J. Bordano. MSC Internal Note No. 66-FM-22
National Aeronautics and Space Administration, Manned Spacecraft Center,
Houston, Texas. April 14, 1966

TABLE OF CONTENTS

Orbital Mechanics

1. Kepler-Newton and the Two Body Problem
2. Concept of a Gravitational Potential
3. Gravitational Attraction with a Distributed Source
4. Near Earth Orbits
5. Impulsive Orbit Transfer
6. Perturbation of Orbital Equations
7. Hyperbolic Trajectories
8. Simple Interplanetary Trajectories
9. Orbit Determination
10. Rendezvous as in Gemini IX

Rocket Propulsion

1. Basic Concepts and Specific Impulse
2. Thermodynamics and Gas Dynamics
3. Energy Generation in Rockets
4. Operation of Rocket Systems

ORBITAL MECHANICS

1. Kepler - Newton Deductions Regarding the Two Body Problem

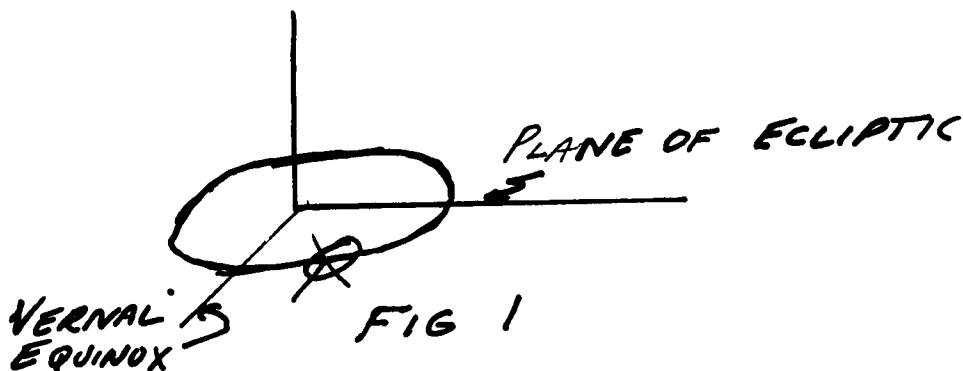
The subject of orbital mechanics has a long history and owes much to the observations of Kepler and of the analysis of Newton.

Kepler formulated three laws which were observed to be obeyed by the planets of the sun. They were:

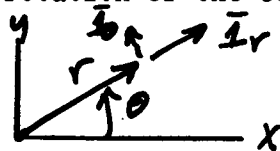
1. The orbits of the planets are ellipses with the sun as the focus.
2. The rate at which area is swept out by the line joining the sun to the planet is constant.
3. The period of the orbit of the planet is proportional to the three halves power of the semi-major axis of the orbit.

These three laws, combined with the Newtonian mechanics were essential in the understanding of classical orbital mechanics.

One of the first questions which might be asked is: In what coordinate system are these observations valid? This is not a trivial question and will lead to some thought in defining an "inertial frame of reference." We will simply define an inertial frame of reference as one wherein Newton's equations of motion are assumed to hold, with the assumption that relativistic effects can be ignored. The coordinate system in which the Kepler laws will be considered is called a "Heliocentric Coordinate System." This system is sun centered, has one coordinate plane in the plane of the earth's axis (plane of the ecliptic), and has a particular direction in this plane defined by the intersection of the earth's equatorial plane with the ecliptic. This direction is called the vernal equinox.



In examining Kepler's laws, one notes first that the statement of the first law implies that the orbit is planar, and that the sun plays a rather central role. Consequently, we shall examine the form of Newton's equations in plane polar coordinates. These equations may be derived simply by the so called "generalized coordinates," but we shall derive them in a more elementary form, since the interpretation of the components will be more intuitive.



The equations of transfer are: $x = r \cos \theta$, $y = r \sin \theta$

In the inertial frame of reference, a vector can be decomposed in many ways. For example, a vector \vec{U} can be written as:

$$\vec{U} = U_x \vec{i}_x + U_y \vec{i}_y = U_r \vec{i}_r + U_\theta \vec{i}_\theta \quad (1.1)$$

Where \vec{i}_x , \vec{i}_y are unit vectors in the direction of increasing x and y , and

\vec{i}_r , \vec{i}_θ are unit vectors in the direction of increasing r and θ . U_r and U_θ

are related to U_x and U_y through

$$U_r = U_x \cos \theta + U_y \sin \theta \quad (1.2)$$

$$U_\theta = -U_x \sin \theta + U_y \cos \theta$$

The velocity vector

$$\vec{V} = \dot{x} \vec{i}_x + \dot{y} \vec{i}_y = (\dot{r} \cos \theta - r \sin \theta \dot{\theta}) \vec{i}_r + (\dot{r} \sin \theta + r \cos \theta \dot{\theta}) \vec{i}_\theta \quad (1.3)$$

$$\vec{V} = \dot{r} \vec{i}_r + r \dot{\theta} \vec{i}_\theta$$

\dot{r} and $r\dot{\theta}$ then represent the instantaneous decomposition of the velocity into the radial and circumferential directions.

Similarly the acceleration:

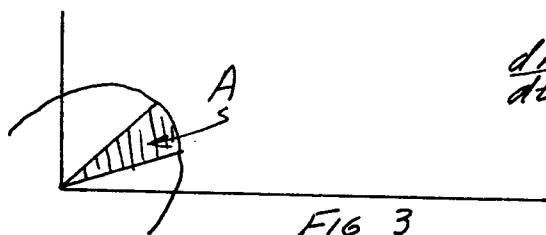
$$\begin{aligned}\bar{a} &= \ddot{x} \bar{i}_x + \ddot{y} \bar{i}_y \\ &= (\ddot{r} \cos \theta - 2\dot{r} \sin \theta \dot{\theta} - r \cos \theta \dot{\theta}^2 - r \sin \theta \ddot{\theta}) \bar{i}_x \\ &\quad + (\ddot{r} \sin \theta + 2\dot{r} \cos \theta \dot{\theta} - r \sin \theta \dot{\theta}^2 + r \cos \theta \ddot{\theta}) \bar{i}_y \\ &= (\ddot{r} - r\dot{\theta}^2) \bar{i}_r + (r\ddot{\theta} + 2\dot{r}\dot{\theta}) \bar{i}_\theta \\ &= a_r \bar{i}_r + a_\theta \bar{i}_\theta\end{aligned}\tag{1.4}$$

If $\dot{\theta}$ is constant, we refer to the term $-r\dot{\theta}^2$ in a_r as the centripetal acceleration, and to the term $2\dot{r}\dot{\theta}$ in a_θ as the Coriolis acceleration.

Note that

$$a_\theta = r\ddot{\theta} + 2\dot{r}\dot{\theta} = \frac{1}{r} \frac{d}{dt} (r^2 \dot{\theta})\tag{1.5}$$

$r^2 \dot{\theta}$ can be identified as the magnitude of the specific angular momentum of the particle about the origin. It also has another interpretation. Consider the area swept out by the radius vector. A will be a function of time and:



$$\frac{dA}{dt} = \frac{1}{2} r v_\theta = \frac{1}{2} r^2 \dot{\theta}\tag{1.6}$$

Let us apply these results to the orbital mechanics problem. For a given planet in the plane of its orbit we say: $m\bar{a} = \bar{F}$

Where \bar{F} is the applied force. This equation can be decomposed into its radial and tangential components to yield:

$$\begin{aligned}m(\ddot{r} - r\dot{\theta}^2) &= F_r \\ \frac{m}{r} \frac{d}{dt} (r^2 \dot{\theta}) &= F_\theta\end{aligned}\tag{1.7}$$

We recall Kepler's second law $\frac{d}{dt} \left(\frac{dA}{dt} \right) = 0$
 or $\frac{d}{dt} (k r^2 \dot{\theta}) = 0$

(1.8)

From this we deduce the property that the applied force on the planet has no component in the tangential direction, since from equation (1.8), $F_{\theta} = 0$

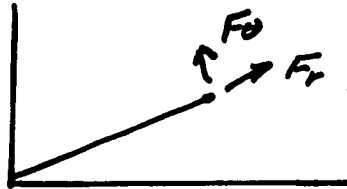


FIG 4

Kepler's second law then implies that the force which attracts the planet to the sun is a Central force. That is, it is a force which is directed towards the sun. For a given orbit of this type we will define:

$$r^2 \dot{\theta} = p = \text{const.} \quad (1.9)$$

We are left with the remaining equation

$$m (\ddot{r} - r \dot{\theta}^2) = F_r \quad (1.10)$$

And the knowledge that the orbits are ellipses, with the sun at the focus.

Let us review a part of analytic geometry with the idea of defining an ellipse in polar coordinates. Recall that an ellipse can be defined as the locus of points in a plane such that the ratio of the distance from a given point (focus) to the distance from a given line (direction) is equal to a constant $e < 1$. e is called the eccentricity.

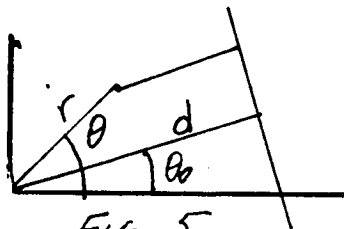


FIG 5.

This may be represented in the form, $\frac{r}{d} = \frac{r}{l - r \cos(\theta - \theta_0)} = e \quad (1.11)$

Where the notation is defined in figure 5

$$\text{OR: } r = \frac{le}{1 + e \cos(\theta - \theta_0)} \quad (1.12)$$

Recall a different form for the definition of an ellipse.

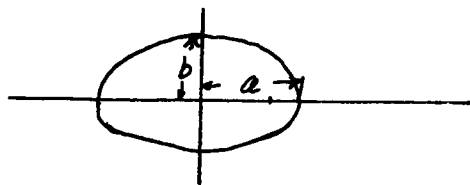


Fig 6.

$$\left. \begin{aligned} \frac{x^2}{a^2} + \frac{y^2}{b^2} &= 1 \\ \text{(Foci at } (\pm c, 0)) \\ c &= \sqrt{a^2 - b^2} \\ e &= \frac{c}{a} \end{aligned} \right\} 1.13$$

a is called the semi-major axis, and by inspection

$$2a = r_{\max} + r_{\min} = Le \left(\frac{1}{1-e} + \frac{1}{1+e} \right) = \frac{2Le}{1-e^2} \quad 1.14$$

Hence we may write $Le = a(1-e^2)$, and we shall use as our standard form.

$$r = \frac{a(1-e^2)}{1+e \cos(\theta-\theta_0)} \quad , \quad \begin{aligned} r_{\max} &= a(1+e) \text{ apogee} \\ r_{\min} &= a(1-e) \text{ perigee} \end{aligned} \quad \left. \right\} 1.15$$

Now $\dot{r} = \frac{a(1-e^2)}{1+e \cos(\theta-\theta_0)} \cdot e \sin(\theta-\theta_0) \dot{\theta} = \frac{pe \sin(\theta-\theta_0)}{a(1-e^2)} \quad 1.16$

Then $\ddot{r} = \frac{pe \cos(\theta-\theta_0) \dot{\theta}}{a(1-e^2)} = \frac{p^2}{r^2} \frac{e \cos(\theta-\theta_0)}{a(1-e^2)} \quad 1.17$

$$m(\ddot{r} - r\dot{\theta}^2) = \frac{mp^2}{r^2} \left(\frac{e \cos(\theta-\theta_0)}{a(1-e^2)} - \frac{1}{r} \right) = -\frac{p^2}{r^2} \frac{m}{a(1-e^2)} \quad 1.18$$

We have established that, at least for a particular orbit, the applied force is proportional to $1/r^2$ and is directed towards the sun.

Consider now the third law

$$T = Ka^{3/2}$$

Where T is the period of the orbit, and K is a constant of proportionality such that equation (1.9) holds true for the planets of the sun. We can compute the period by noting that the rate at which area is swept out by the radius vector is constant and equal to $\frac{P}{2}$.

The total area of the ellipse will be swept out in one period. Hence,

$$\frac{1}{2} PT = \pi ab = \pi \sqrt{1-e^2} \cdot a^2$$

Then

$$P = \frac{2\pi}{K} \sqrt{1-e^2} a^{3/2}$$

and the right hand side of equation (1.18) becomes

$$-\frac{p^2}{r^2} \frac{m}{a(1-e^2)} = -\frac{4\pi^2 m}{K^2 r^2} = Fr \quad 1.20$$

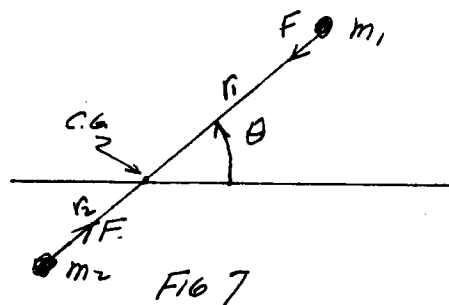
The attractive force of the sun on a planet is then proportional to the mass of the planet, inversely proportional to r^2 , and the constant of proportionality is independent of the orbit chosen.

Newton postulated a law of gravitational attraction such that the two masses m_1 and m_2 attract each other mutually with a force of magnitude $\frac{r m_1 m_2}{r^2}$, where r is a universal gravitational constant. For the sun, with the previously determined value of K (from observation)

$$r m_s = \frac{a T^2}{K^2}, \quad p = \sqrt{r m_s a (1 - \epsilon^2)} \quad 1.21$$

Consider now, application to a system in which m_1 and m_2 are not sufficiently different in magnitude that one mass can be considered to be standing still.

Introduce a polar coordinate system with center at the center of gravity or Barycenter of the two body system, as in figure 7.



$$\left. \begin{aligned} m_1 r_1 &= m_2 r_2 \\ r_1^2 \ddot{\theta} &= \text{CONST} \\ r_2^2 \ddot{\theta} &= \text{CONST} \\ r &= r_1 + r_2 \end{aligned} \right\} 1.22$$

Hence, we can write

$$\ddot{r}_1 - r_1 \dot{\theta}^2 = - \frac{r m_2}{r^2} \quad 1.23$$

$$\begin{aligned} \text{or } \ddot{r} - r \dot{\theta}^2 &= - \frac{r (m_1 + m_2)}{r^2} \\ \text{and } r^2 \dot{\theta} &= \text{const} \end{aligned} \quad \left. \right\} 1.24$$

Now, equation (1.24) is identical in form to the previously studied restricted problem, but the attractive mass has been replaced by the sum of the two masses, and r is not measured from a stationary point in inertial

2. Concept of a Gravitational Potential

A force field is said to be conservative if the work done in opposing this field around a closed contour is zero. In cartesian coordinates it:

$$\mathbf{F} = F_x \mathbf{i}_x + F_y \mathbf{i}_y + F_z \mathbf{i}_z \quad (2.1)$$

Then the work done against this field is:

$$W = - \int_{P_1}^{P_2} F_x dx + F_y dy + F_z dz \quad (2.2)$$

Where P_1, P_2 represent the starting and end points of the integration and C represents the contour over which the integration is performed. We shall not devote much time to the subject of line integrals, but will simply note in passing that if $P_1 = P_2$ implies $W=0$, regardless of C then the field is conservative. If F_x, F_y and F_z are continuously differentiable, one may show that, with certain restrictions on the domain of validity, an irrotational stationary field $(\nabla \times \mathbf{F} = \frac{\partial \mathbf{F}}{\partial t} = 0)$ is conservative and vice versa. Furthermore, it can be represented as the gradient of a scalar $(-\phi)$. ϕ will be called the potential function which generates \mathbf{F} .

Then
$$\mathbf{F} = - \text{grad } \phi = - \frac{\partial \phi}{\partial x} \mathbf{i}_x - \frac{\partial \phi}{\partial y} \mathbf{i}_y - \frac{\partial \phi}{\partial z} \mathbf{i}_z \quad (2.3)$$

Consider the motion of a particle which accelerates under the influence of a conservative force field \mathbf{F} .

$$\left. \begin{aligned} m \ddot{x} &= F_x = - \frac{\partial \phi}{\partial x} \\ m \ddot{y} &= F_y = - \frac{\partial \phi}{\partial y} \\ m \ddot{z} &= F_z = - \frac{\partial \phi}{\partial z} \end{aligned} \right\} \quad (2.4)$$

If we multiply these equations by \dot{x}, \dot{y} and \dot{z} respectively, we obtain:

$$m(\dot{x}\ddot{x} + \dot{y}\ddot{y} + \dot{z}\ddot{z}) = - \left(\frac{\partial \phi}{\partial x} \dot{x} + \frac{\partial \phi}{\partial y} \dot{y} + \frac{\partial \phi}{\partial z} \dot{z} \right) \quad (2.5)$$

and, with ϕ independent of time

$$\left. \begin{aligned} \frac{d}{dt} \left(\frac{1}{2} m (\dot{x}^2 + \dot{y}^2 + \dot{z}^2) + \phi \right) &= 0 \\ \therefore \frac{1}{2} m (\dot{x}^2 + \dot{y}^2 + \dot{z}^2) + \phi &= \text{const} \end{aligned} \right\} \quad (2.6)$$

(Note that $\frac{dy}{dt} = \frac{\partial y}{\partial x} \frac{dx}{dt} + \frac{\partial y}{\partial y} \frac{dy}{dt} + \frac{\partial y}{\partial z} \frac{dz}{dt}$)

Equation (2.6) is called the energy integral for a conservative system.

The two terms on the left side are the Kinetic and Potential energy respectively.

Let us return to the system used in studying the motion of the planets about the sun. With a sun centered cartesian system, the force on a mass m which is located at (x, y, z) is given by

$$\left. \begin{aligned} F_x &= - \frac{r m_s m x}{r^3} \\ F_y &= - \frac{r m_s m y}{r^3} \\ F_z &= - \frac{r m_s m z}{r^3} \end{aligned} \right\} \quad 2.7$$

$$(r^2 = x^2 + y^2 + z^2)$$

and it may be easily shown that this is a conservative system where

$$\psi = - \frac{r m_s m}{r^{1/2}} \quad 2.8$$

Consequently, in the previously discussed orbits, we should be able to

$$\text{show that } \frac{1}{2} m (\dot{x}^2 + \dot{y}^2 + \dot{z}^2) = \text{const} + \frac{r m_s m}{r} \quad 2.9$$

In the plane polar coordinate system used before,

$$\dot{x}^2 + \dot{y}^2 + \dot{z}^2 = \dot{r}^2 + r^2 \dot{\theta}^2 \quad 2.10$$

But

$$\dot{r}^2 + r^2 \dot{\theta}^2 = \frac{p^2 e^2 \sin^2(\theta - \theta_0)}{a^2 (1 - e^2)} + \frac{p^2}{r^2} \quad 2.11$$

$$= \frac{p^2}{a^2 (1 - e^2)} [(1 + e^2) + 2e \cos(\theta - \theta_0)]$$

$$\text{or } v^2 = r m_s \left[\frac{2}{r} - \frac{1}{a} \right] \quad 2.12$$

Equation (2.12) is called the energy or Vis Viva integral for the restricted two body problem.

Certain facts are obvious.

The maximum radius will yield the minimum speed.

$$r_{\text{apogee}} = a(1+\epsilon), \quad r_{\text{perigee}} = a(1-\epsilon)$$

Then

$$v_{\text{min}}^2 = \frac{r m_s}{a} \left(\frac{1-\epsilon}{1+\epsilon} \right)$$

And

$$v_{\text{max}}^2 = \frac{r m_s}{a} \left(\frac{1+\epsilon}{1-\epsilon} \right)$$

For a circular orbit $r=a$ and:

$$v_c^2 = \frac{r m_s}{r}$$

2.14

For a parabolic orbit, $a \rightarrow \infty$

$$v_p^2 = \frac{2 r m_s}{r}$$

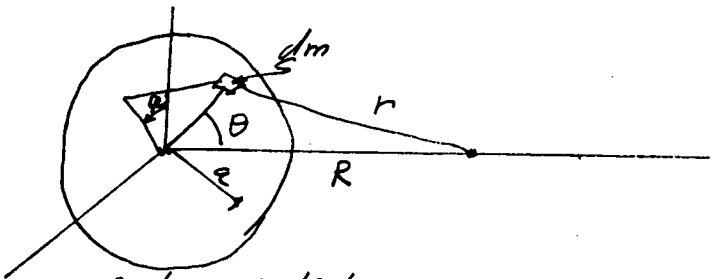
2.15

And the speed required to leave the gravitational system if one is at a given radius r is equal to the square root of two times the speed required to keep it in a circular orbit at that radius.

3. Gravitational Attraction With Distributed Mass

If the mass is not a point mass, but is distributed over a volume, one uses a further postulate of Newton's; that the attractive force of a sum of masses is equal to the vector sum of their attractive forces. Since the operation of taking a gradient is linear, we can either superimpose the forces or their potentials.

Consider now the potential due to a spherical shell. The potential due to an elementary mass is simply $-\frac{\gamma dm}{r}$. We then evaluate the potential as indicated in Figure 8.



Now $dm = \rho a^2 da \sin \theta d\theta d\phi$, where θ and ϕ are spherical coordinates of a point on the surface, a is the radius of the spherical shell, and

da is the thickness. ρ is the density of the material. By the law of Cosines, $r = \sqrt{R^2 + a^2 - 2aR \cos \theta}$ 3.1

$$\text{and } -\varphi = \gamma \rho a^2 da \int_0^{2\pi} \int_0^\pi \frac{\sin \theta d\theta d\omega}{\sqrt{R^2 + a^2 - 2aR \cos \theta}} = \gamma \rho a^2 da \frac{2\pi}{aR} [(R+a) - |R-a|] \quad 3.2$$

where the absolute value is necessary to guarantee a continuous integral between $\theta = 0$ and $\theta = \pi$

$$- \varphi = \frac{4\pi \rho a^2 da \cdot \gamma}{R}, \quad R > a.$$

$$- \varphi = 4\pi \rho a da \cdot \gamma, \quad R < a.$$

Consequently, we see that, outside the sphere, the potential is exactly the form of a point mass potential with mass $4\pi \rho a^2 da$ which is thus equal to the mass of the shell, and which is located at the center of the shell.

On the other hand, inside the shell, the potential is independent of position, and the resulting force will then be zero.

From this we can deduce that, outside a spherical body with spherical symmetry, the attraction is identical to that of a point mass with the same value as the mass of the sphere, located at the center of the sphere.

Unfortunately, the earth is not exactly a sphere, but it is quite close to being spherical. However, to first order, and with reasonable accuracy, we can approximate the earth's gravitational field as being spherical. The deviations may be approximated by an expansion in terms of spherical harmonics.

To first order, then, for near earth trajectories we will consider the earth to be approximately spherical. For purposes of calculation, we will use the following values for earth parameters.

$$\gamma m_e = 4 \times 10^{14} \text{ M}^3/\text{sec}^2$$

$$R_e = 6370 \text{ Km.}$$

Let us consider some simple questions. The ratio of moon to earth mass is .0122, and its radius is 1738. What is one lunar g?

$$\frac{\text{lunar } g}{\text{earth } g} = \frac{R_e^2}{R_m^2} \cdot \frac{m_m}{m_e} = .0122 \cdot (3.67)^2 = .163$$

If the mean distance from earth to moon center is 60.3 earth radii, what is the distance from the earth center to the Barycenter?

$$r_1 = .0122 r_2 = .0122 (r - r_1)$$

$$r_1 = .021 r = .021 \times 60.3 r_e = .73 r_e.$$

If the orbital period is 27.3 days, what is r_{me} ?

$$T = \frac{2\pi}{\sqrt{r_{me}}} r^{3/2}$$

$$27.3 \times 24 \times 3600 = \frac{2\pi}{\sqrt{r_{me}}} (60.3 \times 6.37 \times 10^6)^{3/2}.$$

$$\sqrt{r_{me}} = 2 \times 10^7 \left(\frac{M^{3/4}}{Sec} \right).$$

But we already know that one earth $g = 9.8$ meters per second, and

$$9.8 = \frac{r_{me}}{(6.37 \times 10^6)^2}$$

$$r_{me} = 4 \times 10^{14}$$

so that the prediction based on moon period and surface acceleration are the same.

4. Near Earth Trajectories

In the idealized near earth free flight spacecraft trajectories we will then use:

$$\left. \begin{aligned} r &= \frac{a(1-e^2)}{1+e \cos(\theta-\theta_0)} \\ r^2 \dot{\theta} &= p = \sqrt{r_{me} a(1-e^2)} \\ \ddot{r} - r \dot{\theta}^2 &= -\frac{r_{me}}{r^2} \\ v^2 = \dot{r}^2 + r^2 \dot{\theta}^2 &= r_{me} \left(\frac{2}{r} - \frac{1}{a} \right) \\ \dot{r} &= \frac{p e \sin(\theta-\theta_0)}{a(1-e^2)} = \sqrt{r_{me}} \frac{e \sin(\theta-\theta_0)}{\sqrt{a(1-e^2)}} \end{aligned} \right\} 4.1$$

With the previously defined parameters, we note that, near the earth's surface, the circular orbital speed is given by:

$$v_c^2 = \frac{r_{me}}{r} = \frac{4 \times 10^{14}}{6.36 \times 10^6} = 6.27 \times 10^7 \frac{M^2}{Sec^2}$$

$$v_c = 7.9 \times 10^3 \frac{M}{Sec.}$$

4.2

The escape, or parabolic speed, is given by

$$v_p = \sqrt{2} v_c = 1.1 \times 10^4 M/Sec \sim 7 \text{ miles/sec.} \quad 4.3$$

The variation of θ with time, or its inverse may be obtained by noting that:

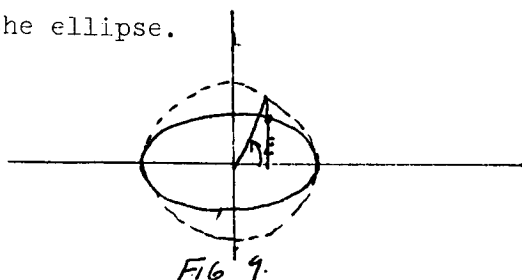
$$p = r^2 \dot{\theta} = \frac{a^2(1-e^2)^2}{[1+e \cos(\theta-\theta_0)]^2} \frac{d\theta}{dt}$$

or

$$dt = \frac{a^2(1-e^2)^2}{p [1+e \cos(\theta-\theta_0)]^2} d\theta$$

} 4.4

The integration may be performed directly, but we shall anticipate its result in order to interpret the time dependence. In so doing, we define an angle called the eccentric anomaly. Return to the cartesian definitions of the ellipse.



$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \leftrightarrow \begin{cases} x = a \cos E \\ y = b \sin E = \sqrt{1-e^2} a \sin E \end{cases} \quad 4.5$$

$$r = \sqrt{(x-ae)^2 + y^2} = \sqrt{a^2(\cos E - e)^2 + a^2(1-e^2)\sin^2 E}$$

$$= a(1 - e \cos E) \quad 4.6$$

Then

$$(1 - e \cos E) = (1 - e^2) / [1 + e \cos(\theta - \theta_0)] \quad 4.7$$

And

$$\sin E = \sqrt{1-e^2} \frac{\sin(\theta - \theta_0)}{1 + e \cos(\theta - \theta_0)} \quad 4.8$$

By differentiating (4.7) we find

$$\sin E dE = \frac{(1-e^2) \sin(\theta - \theta_0) d\theta}{[1 + e \cos(\theta - \theta_0)]^2} \quad 4.9$$

$$dE = \frac{\sqrt{1-e^2} d\theta}{1 + e \cos(\theta - \theta_0)} \quad 4.10$$

If then are combined with (4.4), we find:

$$dt = \frac{a^2(1-e^2)^2}{p} \cdot \frac{1}{(1-e^2)^{3/2} (1 - e \cos E)} dE \quad 4.11$$

$$n \cdot dt = \frac{a^{3/2}}{\sqrt{\mu_{me}}} (1 - e \cos E) dE \quad 4.12$$

and

$$t - t_0 = \frac{a^{3/2}}{\sqrt{\mu_{me}}} (E - e \sin E) \quad 4.13$$

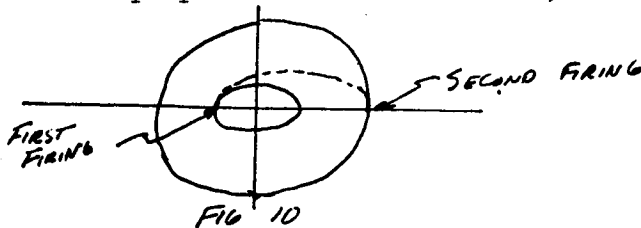
where t_0 is defined such that $E = 0$ at $t = t_0$

In one period, $E_1 \rightarrow E_0 = 2\pi$ and $t_1 - t_0 = \frac{2\pi a^{3/2}}{\sqrt{\mu_{me}}}$. This is consistent with equation (1.24). We will apply this result subsequently. Note: A twenty-four hour orbit would have an a of $\left(\frac{24 \times 3600}{2\pi} \times 2 \times 10^7\right)^{2/3} = 4.24 \times 10^7 \text{ M.}$

Question, do we really want it to be a twenty-four hour satellite?

5. Impulsive Orbit Transfer in a Plane

A. Transfer Ellipse- Suppose that we are in a "parking orbit" which is nearly circular (perigee= 6700 Km, apogee= 7200 Km), and we wish to establish a twenty-four hour circular orbit by two impulsive thrustings, how can we perform the maneuver? We could conceivably fire a rocket anywhere in the parking orbit, coast until proper altitude is reached, then fire again.



We shall assume that the first firing will take place at either apogee or perigee of the parking orbit, and that the newly established "transfer" orbit will have its apogee at the altitude of a twenty-four hour circular orbit. A transfer from a circular to a circular orbit is called a Hohman transfer maneuver.

First, define the parking ellipse and its characteristics.

$$2a = r_{max} + r_{min} = 6700 + 7200 = 13,900 \text{ Km.}$$

$$r_{max} = 7200 = a(1+e) = 6,950(1+e)$$

$$e = .036$$

$$v_{perigee}^2 = \gamma_{mc} \left(\frac{2}{r_{min}} - \frac{1}{a} \right) = 6.18 \times 10^7 \frac{M^2}{Sec^2} : v_{perigee} = 7.81 \times 10^3 \frac{M}{Sec}$$

$$v_{apogee}^2 = \gamma_{mc} \left(\frac{2}{r_{max}} - \frac{1}{a} \right) = 5.36 \times 10^7 \frac{M^2}{Sec^2} : v_{apogee} = 7.28 \times 10^3 \frac{M}{Sec}$$

We will consider then two possibilities for a transfer orbit. For both orbits, the apogee will be $4.24 \times 10^7 M$. The two perigee radii will be 6700 and 7200 Km.

The following are characteristics of the two transfer orbits.

	a	$v_{perigee}$	v_{apogee}
Park. Perigee	$2.455 \times 10^7 M$	$1.016 \times 10^4 \frac{M}{Sec}$	$1.604 \times 10^3 \frac{M}{Sec}$
Park apogee	$2.48 \times 10^7 M$	$.976 \times 10^4 \frac{M}{Sec}$	$1.655 \times 10^3 \frac{M}{Sec}$

The speed in the twenty-four hour circular orbit will be $3.07 \times 10^3 \frac{M}{Sec}$.

For the first procedure then, a velocity increment of $(1.016 - .781) \times 10^4 = 2.35 \times 10^3 \frac{M}{Sec}$ will be required at transfer apogee.

For the second procedure, an incremental speed $(.976 - .728) \times 10^4 = 2.48 \times 10^3 \frac{M}{Sec}$ will be required at transfer perigee and $(3.07 - 1.655) \times 10^3 = 1.415 \times 10^3 \frac{M}{Sec}$ will be required at transfer apogee.

General equations for requirements for transfer from a circular orbit to another circular orbit of different altitude are readily obtained. If the subscript 1 and 2 refer to the two circular orbits and if the subscript t designates the transfer orbit.

$$\left. \begin{aligned} V_1 &= \sqrt{\frac{r m_e}{r_1}} \quad , \quad V_2 = \sqrt{\frac{r m_e}{r_2}} \\ V_{t_1} &= \sqrt{r m_e \left(\frac{2}{r_1} - \frac{2}{r_1 + r_2} \right)} \quad , \quad V_{t_2} = \sqrt{r m_e \left(\frac{2}{r_2} - \frac{2}{r_1 + r_2} \right)} \end{aligned} \right\} 5.1$$

and, if $r_2 > r_1$

$$\left. \begin{aligned} \Delta V_1 &= V_{t_1} - V_1 = \sqrt{\frac{r m_e}{r_1}} \left(\sqrt{2 - \frac{2}{1 + \frac{r_2}{r_1}}} - 1 \right) \\ \Delta V_2 &= \sqrt{\frac{r m_e}{r_2}} \left(1 - \sqrt{2 - \frac{2}{1 + \frac{r_1}{r_2}}} \right) \end{aligned} \right\} 5.2$$

6. Perturbation of the Orbital Equations

A perturbation to a nominal condition or state is simply a deviation from this nominal. If this deviation is slight, the perturbation can be represented as a linear function or functional of the cause of the deviation. The deviation of an orbit from its nominal trajectory may be caused by disturbing forces (i.e. radiation pressure, gravitational anomalies, small thrust devices) or it may be caused by improper initial or injection conditions.

Let us consider the effect on a given orbit of firing a very low impulse device with the impulse aligned with the velocity vector, and fired at perigee. Suppose the impulse device delivers an increment at speed ΔV which is small compared to the local orbital speed.

We know that

$$V^2 = r m_e \left(\frac{2}{r} - \frac{1}{a} \right) \quad 6.1$$

After firing, the speed is changed suddenly from a value V_1 to $V_1 + \Delta V$. If it is applied instantaneously, $\frac{2}{r}$ cannot be changed during the firing, but a will be changed. With a small percentage change in V , it would be expected that there would be a small percentage change in a . We shall call the new semi-major axis $a + \Delta a$.

Then we have, just prior to firing

$$v_i^2 = r m_e \left(\frac{2}{r} - \frac{1}{a} \right) \quad 6.2$$

and just after firing

$$(v_i + \Delta v)^2 = r m_e \left(\frac{2}{r} - \frac{1}{a + \Delta a} \right) \quad 6.3$$

We can expand both sides of the equation in the formula,

$$v_i^2 + 2v_i \Delta v + \dots = r m_e \left(\frac{2}{r} - \frac{1}{a} \left\{ 1 - \frac{\Delta a}{a} + \left(\frac{\Delta a}{a} \right)^2 - \dots \right\} \right) \quad 6.4$$

and, to first order, if we subtract equation (6.2) from (6.4), we find,

$$2v_i \Delta v = r m_e \left(\frac{\Delta a}{a^2} \right) \quad 6.5$$

So that Δa can be computed as a linear function of Δv . It is often not only easier, but also more accurate to compute changes by perturbation techniques, especially if small differences in large quantities are required.

The foregoing procedure can be made more general if we simply consider the following:

Let $f(v) = g(r, a)$, and let v be changed by a small amount in such a way that

r is unchanged. Then how can we determine the change in a ?

$$f(v + \Delta v) = f(v) + \frac{\partial f}{\partial v} \Delta v + \dots \quad 6.6$$

$$g(r, a + \Delta a) = g(r, a) + \frac{\partial g}{\partial a} \Delta a + \dots \quad 6.7$$

$$\text{Then. } \frac{\partial f}{\partial v} \Delta v = \frac{\partial g}{\partial a} \Delta a. \quad 6.8$$

Let us consider an example. A spacecraft is in a circular orbit of altitude 200 Km. A ball is thrown backwards at a speed of 5 meters per second. The spacecraft makes one revolution. Where is the ball relative to the spacecraft?

$$v_i^2 = \frac{r m_e}{r} = \frac{4 \times 10^8}{6.57 \times 10^6} = .609 \times 10^8$$

$$v_i = 7.76 \times 10^3 \text{ m/sec.}$$

Equation (6.5) becomes:

$$-2 \times 7.76 \times 10^3 \times 5 = 4 \times 10^{14} \left\{ \frac{\Delta a}{(6.57 \times 10^6)^2} \right\}$$

$$\Delta a = -8.39 \times 10^3 \text{ m.}$$

But recall that $T = \frac{2\pi a^{3/2}}{\sqrt{r m_e}}$

By the same procedure

$$\Delta T = \frac{2\pi \cdot \frac{3}{2} a^{1/2} \Delta a}{\sqrt{r m_e}} = -10.11 \text{ sec.}$$

The ball will reach the release point ten seconds before the vehicle, and will be some fifty miles ahead of the vehicle at that point.

One can find the response of an orbit to an impulse in either the radial or tangential direction applied at θ_1 . If r_1 is the corresponding value of r , the orbit will be changed by changing a, e and θ_0 . One may determine these quantities by considering the variation in the equations for $\dot{r}, r\dot{\theta}$ and r .

For example,

$$\left. \begin{aligned} \Delta \dot{r} &= \sqrt{rme} \left\{ \frac{e \cos(\theta_1 - \theta_0)}{\sqrt{a(1-e^2)}} \Delta \theta_0 - \frac{e \sin(\theta_1 - \theta_0)}{2a^{3/2} \sqrt{1-e^2}} \Delta a + \frac{\Delta e}{(1-e^2)^{3/2}} \sin(\theta_1 - \theta_0) \right\} \\ (\Delta \dot{\theta} r_1) &= \frac{\sqrt{rme}}{r_1} \left\{ \frac{1}{2} \frac{\sqrt{1-e^2}}{a} \Delta a + \sqrt{\frac{a}{1-e^2}} e \Delta e \right\} \\ 0 &= \Delta e \left\{ r_1 \cos(\theta_1 - \theta_0) + 2a e \right\} + \Delta \theta_0 r_1 e \sin(\theta_1 - \theta_0) - (1-e^2) \Delta a. \end{aligned} \right\} 6.9$$

These three equations allow a unique solution for $\Delta e, \Delta \theta_0$ and Δa as linear functions of $\Delta \dot{r}$ and $r \Delta \dot{\theta}$.

Up to this point, we have been concerned only with variation or perturbation of functional relations. Similar perturbations can be made on differential equations, and the resulting equations will be shown to be linear.

We will consider the procedure first as applied to a simple differential equation, then we will make a more serious application.

Suppose that we have a linear spring mass system, which at time $t=0$ has a displacement y_0 and a velocity \dot{y}_0 . Then:

$$\ddot{y} + \omega_0^2 y = 0 \quad \begin{cases} y(0) = y_0 \\ \dot{y}(0) = \dot{y}_0 \end{cases} \quad 6.10.$$

The solution of the system is:

$$y = y_0 \cos \omega_0 t + \frac{\dot{y}_0}{\omega_0} \sin \omega_0 t \quad 6.11$$

This will be called the nominal solution. Examine the question. Suppose the ω_0, y_0 and \dot{y}_0 are not matched accurately. How will the resulting motion be modified? Since we already know the complete solution, we can use the procedures which were just examined, i.e.

$$\Delta y = \Delta y_0 \cos \omega_0 t + \frac{\Delta \dot{y}_0}{\omega_0} \sin \omega_0 t - \left(y_0 t + \frac{\dot{y}_0}{\omega_0} \right) \sin \omega_0 t \Delta \omega_0 + \frac{\dot{y}_0 t}{\omega_0} \cos \omega_0 t \Delta \omega_0. \quad 6.12$$

We could, however, have proceeded differently. We could have studied the generation of and propagation of error.

For example, suppose we consider the system equation and assume that $y = y_{\text{nominal}} + \delta y$.

Here δy is called the "variation" of y . y_{nominal} satisfies the exact system, but y satisfies the modified (in error) system.

$$\text{Then, } \ddot{y} + (\omega_0 + \Delta\omega)^2 y = 0 \quad \left. \begin{array}{l} y(0) = y_0 + \Delta y_0 \\ \dot{y}(0) = \dot{y}_0 + \Delta \dot{y}_0 \end{array} \right\} 6.14$$

$$\text{and } \ddot{y}_{\text{nom}} + \delta \ddot{y} + (\omega_0^2 y_{\text{nom}}) + 2\omega_0 \Delta\omega y_{\text{nom}} + \omega_0^2 \delta y = 0 \quad 6.15$$

$$\text{and } \left. \begin{array}{l} \delta y(0) = \Delta y_0 \\ \delta \dot{y}(0) = \Delta \dot{y}_0 \end{array} \right\} 6.16$$

Only first order terms have been retained in the differential equation. If the nominal equation is subtracted from (6.15), then the equation becomes

$$\frac{1}{\omega_0} (\delta \ddot{y} + \omega_0^2 \delta y) = \Delta\omega_0 y_{\text{nom}} = -\Delta\omega_0 (y_0 \cos \omega_0 t + \frac{\dot{y}_0}{\omega_0} \sin \omega_0 t) \quad 6.17$$

The solution to (6.16) subject to the initial conditions (6.17) is:

$$\delta y = \Delta y_0 \cos \omega_0 t + \frac{\Delta \dot{y}_0}{\omega_0} \sin \omega_0 t - (y_0 t + \frac{\dot{y}_0}{\omega_0}) \sin \omega_0 t \Delta\omega_0 + \frac{\dot{y}_0 t}{\omega_0} \cos \omega_0 t \Delta\omega_0 \quad 6.18$$

The Δy and δy are seen to be equal. The primary reason for studying the variational form of the differential equation is that usually, no such general solution to a differential equation can be explicitly displayed. The method can be outlined in a very general form.

Suppose,
$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \dots, x_n, t; b_1, b_2, \dots) \\ \dot{x}_2 &= f_2(x_1, \dots, x_n, t; b_1, b_2, \dots) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \dots, x_n, t; b_1, b_2, \dots) \end{aligned} \quad 6.19$$

and suppose a nominal solution corresponds to some set of initial conditions x_1^0, \dots, x_n^0 , and a set of parameters b_1^0, \dots . Let this solution be designated by $\bar{x}_1, \dots, \bar{x}_n$.

Then for deviations in initial conditions or in values of the parameters the perturbations satisfy.

$$\delta \dot{x}_i = \frac{\partial f_i}{\partial x_1} \delta x_1 + \frac{\partial f_i}{\partial x_2} \delta x_2 + \dots + \frac{\partial f_i}{\partial x_n} \delta x_n + \frac{\partial f_i}{\partial b_1} \delta b_1 + \dots \quad 6.20$$

This class of equations is often referred to as "variational equations."

Example- $\ddot{r} - r\dot{\theta}^2 = -\frac{r m_e}{r^3} + b_r$
 $\frac{d}{dt}(r^2 \dot{\theta}) = r b_\theta$ } 6.21

Where b_r and b_θ are disturbing forces with a nominal value of zero.

One technique is to solve the nominal case ($b_r = b_\theta = 0$) and then to consider the disturbing forces to be perturbations, i.e.

$$\begin{aligned} \delta \ddot{r} - \delta r \dot{\theta}^2 - 2r \dot{\theta} \delta \dot{\theta} &= \frac{2r m_e}{r^3} \delta r + b_r \\ \frac{d}{dt}(2r \dot{\theta} \delta r + r^2 \delta \dot{\theta}) &= r b_\theta. \end{aligned} \quad \left. \vphantom{\frac{d}{dt}} \right\} 6.22$$

These two differential equations are linear in δr and $\delta \dot{\theta}$. The nominal equations are presumed to be solved so that the r and $\dot{\theta}$ are known functions of time. Consider now the special case where $b_\theta = 0$, $\delta \dot{\theta} = -\frac{\dot{\theta}}{r} \delta r$. Then:

$$\delta \ddot{r} + \left(\frac{3r \dot{\theta}^2}{r^3} - \frac{2r m_e}{r^3} \right) \delta r = b_r \quad 6.23$$

For the case of a circular orbit this reduces to

$$\delta \ddot{r} + \left(\frac{r m_e}{r^3} \right) \delta r = b_r \quad 6.24$$

or

$$\delta \ddot{r} + \dot{\theta}^2 \delta r = \ddot{b}_r \quad 6.25$$

which is easily solved for any input b_r .

For example, consider the motion of a vehicle with a low thrust engine which is turned on suddenly. Then $b_r = 0, t < 0$, and $b_r = \text{constant}, t > 0$.

The solution is of the form,

$$\delta r = \frac{\ddot{b}_r}{\dot{\theta}^2} (1 - \cos \dot{\theta} t) \quad 6.26$$

Or, examine the effect of aerodynamic drag on orbital decay of a circular orbit. In this case, the aerodynamic drag will induce a negative value of $r b_\theta$.

Then : $2r \dot{\theta} \delta r + r^2 \delta \dot{\theta} = \int_{t_0}^t r b_\theta dt \quad 6.27$

If the drag is nearly constant in magnitude $\int_{t_0}^t r b_\theta dt$ is approximated by $r b_\theta \cdot t$. A solution to the perturbation equation of the formula:

$$\delta r = c_1 t, \quad \delta \dot{\theta} = c_2 t$$

may be found. With $b_r = 0$, equation (6.22) then yields,

$$\begin{aligned} 2r \dot{\theta} c_1 + r^2 c_2 &= r b_\theta \\ -(\dot{\theta}^2 + \frac{2r m_e}{r^3}) c_1 - 2r \dot{\theta} c_2 &= 0 \end{aligned} \quad 6.28$$

But in a circular orbit, $\frac{r m_e}{r^3} = \dot{\theta}^2$,

and we find that

$$C_1 = \frac{2b_0}{\theta}$$

6.29

$$C_2 = -\frac{3b_0}{r}$$

The following interpretation can be given to the foregoing: Recall the

rb_0 is negative. We find that r is changing at a rate C_1 which is $\frac{2b_0}{\theta} < 0$ and that $\dot{\theta}$ is changing at a rate $C_2 = -\frac{3b_0}{r} > 0$. The radius is decreasing, while the angular rate (and incidentally, also the speed) is increasing.

We have then exhibited a solution (not necessarily the correct one) to a set of perturbation equations. If the orbit is decaying, we would expect the aerodynamic drag to increase with time because of the fact that with decreasing altitude the density and speed both increase. However, such an analysis might be valid if the drag changes slowly. For example, if the percentage change in one orbit is small compared to one, this type of analysis could be valid.

A similar result could be achieved by an intuitive procedure.

Note that: $\frac{1}{2} v^2 - \frac{r m_e}{r} = -\frac{1}{2} \frac{r m_e}{a} = E$

6.30

Where E represents the specific kinetic plus potential energy of the orbit.

If we are in a circular orbit, $E = -\frac{1}{2} \frac{r m_e}{r}$. The work done per unit time by the aerodynamic drag is:

$$\frac{dW}{dt} = b r \cdot r \dot{\theta}$$

But, $\frac{dW}{dt} = \frac{dE}{dt} = \frac{1}{2} \frac{r m_e}{r^2} \frac{dr}{dt}$

6.32

and $\frac{dr}{dt} = \frac{2b_0}{\theta}$

7. Hyperbolic Trajectories

For speeds in excess of the local parabolic speeds, trajectories which are no longer elliptical will be required. The general solution to the dynamical equation is in fact:

$$r^2 \dot{\theta} = p$$

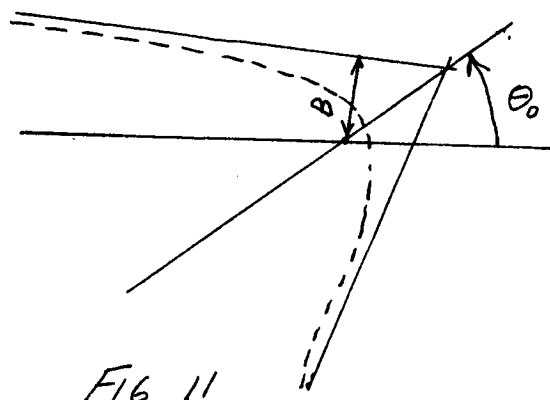
$$r = \frac{a(1-\epsilon^2)}{1 + \epsilon \cos(\theta - \theta_0)}$$

} 7.1

If $\epsilon > 0$, the equation is hyperbolic. We can use the results already tabulated for an ellipse if we allow $\epsilon > 1$ and $a < 0$. Note that $r \rightarrow \infty$ if $\cos(\theta - \theta_0) \rightarrow -\frac{1}{\epsilon}$. These two values of θ correspond to the two asymptotes of the hyperbolia.

The orientation is as indicated in figure 11. The distance B is the distance from the focus is the projection of the velocity at large distances. Now the vis viva integral becomes ;
$$v^2 = \gamma m_e \left(\frac{2}{r} - \frac{1}{a} \right)$$

where a is negative. Note that $-\frac{\gamma m_e}{a}$ may be interpreted as the square of the speed when $r \rightarrow \infty$. We shall designate $v_\infty = \sqrt{-\frac{\gamma m_e}{a}}$. Note that at any point,



$$v^2 = v_p^2 + v_\infty^2 \quad 7.2$$

$$v_p \equiv v(\text{parabolic})$$

If we wish then to launch a vehicle on a journey to Mars, we must supply sufficient speed to escape the earth's gravitational attraction and have sufficient speed left to effect the orbit change from the earth's orbit about the sun to an orbit which would intercept the orbit of Mars.

The quantity B may be evaluated by noting that, at ∞ , p =angular momentum= $v_\infty B$

We consider the question, what hyperbolic orbits will intercept the earth? If we set $r_{\text{perigee}} = r_e$, we find that if
$$B_0 = r_e \left(1 + \frac{v_p^2}{v_\infty^2} \right)^{1/2} \quad 7.3$$

Where v_p is evaluated at the earth's surface, that the trajectory will intercept the earth. If B is less than B_0 , the interception will occur.

B_0 is called a collision length. Note that if $\frac{v_p}{v_0} \gg 1$ (slow approach) then a wide class of trajectories will be captured. If $\frac{v_0}{v_p} \gg 1$ (very fast approach), the collision length is effectively r_e .

8. Interplanetary Trajectories

Suppose we wish to travel from the earth to Jupiter. We must first escape the earth. After escape we must have sufficient residual speed to effect the orbit change on the solar system to take us from the earth to Jupiter. Such a trajectory really needs to be calculated on a computer, but, an approximation to the requirements can be achieved by examining the operation in steps. The procedure is mathematically that of an inner and outer expansion.

We will consider, for this calculation that the planetary orbits are circular, and coplanar.

First, consider the solar system parameters. Jupiter has an orbit radius of about 5.2 times the radius of the earth's orbit. The earth's orbit radius is approximately 1.5×10^{11} m. We will take $r_{ms} = 1.33 \times 10^{20} \frac{M^3}{SEC^2}$.

The earth's speed in orbit is then given by,

$$v_c = \sqrt{\frac{1.33 \times 10^{20}}{1.5 \times 10^{11}}} = 2.98 \times 10^4 \frac{M}{SEC} \quad 8.1$$

A transfer orbit (sun centered) which would have its perigee at the earth's orbit and its apogee at Jupiter's orbit would then be defined by,

$$a = \frac{1+5.2}{2} \times 1.5 \times 10^{11} = 4.65 \times 10^{11} M$$

$$e = \frac{3.15}{4.65} = .678$$

} 8.2

Its velocity at perigee would be given by,

$$v_t^2 = r_{ms} \left(\frac{2}{r_p} - \frac{1}{a} \right)$$

$$v_t = 3.86 \times 10^4 \frac{M}{SEC}$$

In order to effect this interception of Jupiter's orbit then we must escape the earth's gravitational field and have a residual speed which is sufficient to modify the orbit relative to the sun so that the vehicle orbit becomes the transfer orbit. We assume that the gravitational effect of the earth is largely confined to a region of the earth which is small compared to the earth's solar trajectory. With this assumption we decouple the two gravitational effects and say that the trajectory should leave the earth with

$$V_{\infty} = V_t - V_e = (3.86 - 2.98) \times 10^4 = .88 \times 10^4 \frac{M}{Sec} \quad 8.4$$

If the vehicle were fired from a parking orbit, the required speed at burnout would be ;

$$V^2 = V_p^2 + V_{\infty}^2 = \left\{ (1.1 \times 10^4)^2 + (.88 \times 10^4)^2 \right\} = 2.0 \times 10^8 \frac{M^2}{Sec^2} \quad 8.5$$

$$V = 1.41 \times 10^4 \frac{M}{Sec}$$

9. Orbit Determination

In the idealized near earth situation, there is a six parameter family of trajectories which can be defined in geocentric coordinates. If we write the governing equations in the form:

$$\begin{aligned} m\ddot{x} &= - \frac{r m_e m x}{r^3} \\ m\ddot{y} &= - \frac{r m_e m y}{r^3} \\ m\ddot{z} &= - \frac{r m_e m z}{r^3} \end{aligned} \quad \left(r \equiv \sqrt{x^2 + y^2 + z^2} \right) \quad 9.1$$

We are considering three simultaneous second order equations, and six initial conditions are necessary to perform an integration starting at $t=t_0$. If we consider the elliptic trajectories themselves we find five geometrical parameters. The intrinsic properties of the ellipse itself are defined by the two parameters a and e . There are then three parameters required to define the orientation of the ellipse. These are i , the inclination of the orbit,

Ω the angle between the vernal equinox, and the ascending node of the orbit, and θ_0 , the angle between the line of apsides and the ascending node

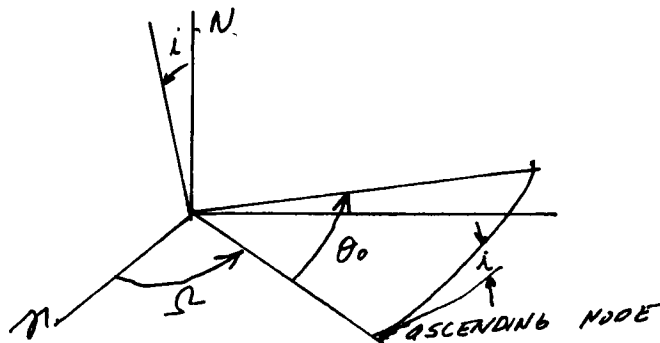


Fig 12.

The sixth parameter is not geometrical but is the time location relation in the orbit itself. If, for example, rendezvous is to be performed, the time is an essential part of the program. How then is orbit determination performed?

First, what data is available?

1. Line of sight measurements. These would be either optical or radar, but are primarily a measurement of local azimuth and elevation. They can, in turn, be transformed to geocentric coordinates if local latitude, longitude, and time are known.
2. Range and angle rate data. These are primarily obtained by a coded pulse and a doppler shift. If the distance of the vehicle from the point of observation is large, this data is usually more accurate than angle data.

What then is an acceptable procedure for orbit determination? If the gravitational attraction were exactly spherically symmetrical inverse square law, and if six measurements were made exactly, these should be sufficient to determine the orbit precisely. However, there are difficulties such as:

1. The gravitational attraction is not ideal. Furthermore, any aerodynamic effect, although small, would cause a perturbation of the orbit.
2. Measurement techniques are inexact. Consequently, the usually accepted procedure for orbit determination is one wherein the governing equations are approximated as carefully as possible, and a highly redundant set of measurements is used to determine the best fit of the assumed form of the trajectory to the observed data.

This technique, with high redundancy can in fact be used to determine more accurately, the form of the gravitational attraction. In general, the technique

is one wherein an estimate is made of the trajectory, further measurements are made and correlated and a revision of the estimate is effected. Further measurement can be used for a next correction, etc.

The problem of proper assessment of data with noise and the processing of this data to effect an orbit determination is a sophisticated exercise in mathematical statistics.

In a simplified form the procedure is one in which the parameters of the problem (say the six initial conditions) are to be determined. One can make an initial estimate, measure the trajectory and define some kind of an error signal. This might be, for example, the output of a time integral of some signal. If for example, a doppler measurement is to be used one can construct a predicted signal and compare it to a measured signal. The error signal could then be defined as some functional of the difference. For example, if the measured signal is $f_1(t)$ and the predicted signal is $f(t)$, we can define

$$\epsilon = \int_{t_0}^t H(\tau) \{f_1(\tau) - f(\tau)\}^2 d\tau \quad 9.2$$

where $H(t)$ is a positive definite weighting function.

Now f is a function of the six initial conditions. The task can then be defined as the determination of the six initial conditions such that ϵ will be minimized. The definition of such items as the best weighting function, and of the best way to continually update the estimate will not be considered here.

Rendezvous Project Gemini IX

1. Gemini Atlas Agena Target Vehicle Launch near circular orbit. 161 nautical mile altitude at an inclination of 28.87° .
 2. Gemini-Titan launch
- 87 N.M. perigee, 146 N.M. apogee at injection. It trails target by 624 N.M.

3. Phase adjustment. $\Delta V = 53.4 \frac{\text{ft}}{\text{sec}}$ positive at first apogee. Raises perigee to 116 N.M. Reduces catchup rate from 6.7°/ orbit to 4.5°/ orbit.
4. Spacecraft correction combined orbit. Phasing-height- out of plane adjustment. Correction (estimate about 5 ft. per sec. required). (trails by 177.8 N.M.) (23 minutes before second apogee).
5. Co-Elliptical maneuver (near second apogee) brings perigee up to apogee altitude, and any remaining out of plane motion is annulled. 52.9 feet per sec. ΔV . Vehicle trails spacecraft by 134 N.M.
6. Switch to rendezvous mode and wait about 62 minutes prior to initiating intercept trajectory.
7. Terminal phase initiation. At range of about 32 N.M., initial a 32.4 feet per second impulse. along 10.5° (27° up). (wt=130°)
8. Twelve minutes after the impulse. (wt=81.8°) an intermediate correction 3 feet per second is applied. After another 12 minutes, a 4 feet per second correction is applied. Here the range is about 4 N.M.
9. Terminal Phase Velocity matching required is 42 feet per second, but is handled by pilot- semi optical technique.

ROCKET PROPULSION

1. Basic Concepts

A rocket is simply a device which expels mass from a vehicle in a given direction with a relative speed so that thrust is achieved.

Consider a vehicle in free space with a mass $m(t)$ which moves in a straight line with speed v , and expels mass backwards at a rate $-\dot{m}(t)$ with an exit speed C . With no external forces acting on the vehicle, the momentum of the vehicle and wake will be independent of time. The rate of change of momentum of the vehicle in the direction of motion is:

$$\frac{d}{dt} (mv) = m\dot{v} + \dot{m}v \quad (1.1)$$

The rate of change of momentum of the wake is simply the rate at which momentum is added through mass addition:

$$-\dot{m}(v-C) \quad 1.2$$

Consequently

$$m\dot{v} + \dot{m}v - \dot{m}(v-C) = 0 \quad 1.3$$

$$\text{or } m\dot{v} = -\dot{m}C = T \quad 1.4$$

The term on the right hand side of 1.4 is often referred to as the vacuum thrust.

In general, the speed C is a function of the propellant and nozzle design, and in many applications, can be assumed to be independent of the burning rate.

With this assumption, equation 1.4 can be integrated to yield:

$$\frac{dv}{C} = -\frac{dm}{m} \quad 1.5$$

$$\text{or } v = v_0 + C \ln \frac{m_0}{m} \quad 1.6$$

Where the subscript 0 designates the initial conditions. High speeds can in general be achieved through high exit speeds C and large ratios of initial to burnout mass.

We will find the concept of impulse to be useful. Impulse is defined as the time integral of a force and has the dimensions of lb sec.

Recall that in one dimensional Newtonian mechanics:

$$F = m \frac{d^2x}{dt^2} \quad 1.7$$

where

$$\begin{aligned} F &= \text{FORCE, (lbs)} \\ m &= \text{MASS (slugs} = \frac{\text{lb sec}^2}{\text{ft}}) \\ x &= \text{DISPLACEMENT (ft)} \end{aligned}$$

so that

$$m \left(\frac{dx}{dt} - \frac{dx}{dt} \Big|_0 \right) = \int_{t_0}^t F dt \quad 1.8$$

And the impulse is a measure of the change in momentum.

In rocket engines, the Specific Impulse, I_{sp} , is defined as the impulse delivered divided of the weight. i.e.

$$I_{sp} = \frac{\int_{t_0}^t T dt}{\text{Wgt. of PROPELLANT}} \quad 1.9$$

From equation (1.8) $\int_{t_0}^t T dt = \int_{t_0}^t -\dot{m} c d\tau = c(m_0 - m_f) = c m_p \quad 1.10$

consequently:

$$I_{sp} = \frac{c m_p}{W_p} = \frac{c}{g} \quad 1.11$$

Where g is the acceleration of gravity at the surface of the earth.

By definition I_{sp} has the dimension of seconds.

It should be emphasized that this is the zero back pressure value which has been used up to this time.

The rocket designer is, among other things, charged with obtaining a design for which the corresponding I_{sp} is large. A good typical value for a chemical propellant rocket engine is 300 sec. We will examine the thermodynamic base of this statement subsequently.

Calculation: Suppose a rocket engine has an I_{sp} of 300 sec., and a velocity change of 25,000 ft/sec is required, what is the minimum mass ratio attainable?

From 1.6. $\ln \frac{m_0}{m} = \frac{25,000}{9,660}, \quad \frac{m_0}{m} = 13.3$

With this class of specific impulses, it is obvious that a very large mass will be required to inject a significant mass into orbit.

2. Thermodynamics and Gas Dynamics

We will consider the quasi-one dimensional steady form of the gas dynamic equations. These equations will be the continuity, momentum, and energy equations. We shall consider only the inviscid form of the equations, and not consider dissipative effects. Much of the qualitative and approximate quantitative character of the flow may be obtained from these equations. We consider a gas flow of an idealized gas with a streamtube area A which is a function of distance x . The velocity u , temperature T , and density ρ will be needed to characterize the flow.

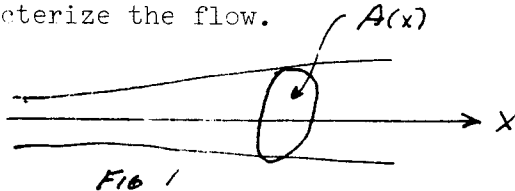


Fig 1

The continuity equation becomes

$$\rho u A = \text{CONST.} \quad 2.1$$

The momentum equation may be obtained rather easily by Newtonian force-momentum balance in three dimensions and will be found in any book in fluid mechanics.

In one dimensional form, the conservation of momentum equation becomes:

$$\rho u \frac{\partial u}{\partial x} = - \frac{\partial p}{\partial x} \quad 2.2$$

In order to discuss the energy equation we recall the form of the first law of thermodynamics which states:

$$dq = d\bar{U} + p d\bar{V}$$

Where $d\bar{q}$ is the heat added, \bar{U} is the internal energy, p is the pressure and \bar{V} is the volume. $\bar{h} = \bar{U} + p\bar{V}$ is the enthalpy of the gas. If we take the form of equation (2.3) which refers to a unit mass, we obtain:

$$dg = dU + P d\left(\frac{1}{\rho}\right) = dh - \frac{1}{\rho} d\rho \quad (2.4)$$

Where the unbarred variables refer to the specific value, or value per unit mass.

For a perfect gas, recall that

$$P = \rho R T$$

(R is universal gas constant divided by the molecular weight)

$$dg/\rho = C_v dT = dU; \quad U = C_v T$$

2.5

$$dg/\rho = C_p dT = dh; \quad h = C_p T$$

But

$$h = U + \frac{P}{\rho} = U + RT$$

2.6

$$C_p = C_v + R$$

$$R = C_p - C_v$$

2.7

Also, recall that the entropy S is a property of state and is defined through-

$$ds = \frac{dg}{T}$$

2.8

For a perfect gas

$$\begin{aligned} ds &= C_v \frac{dT}{T} + \frac{P}{T} d\left(\frac{1}{\rho}\right) \\ &= C_v d \log T + R d \log \left(\frac{1}{\rho}\right) = C_v d \log \frac{T}{\rho^{r-1}} = C_v d \log \frac{P}{\rho^r} \end{aligned} \quad (2.9)$$

And we define the ratio $\frac{C_p}{C_v} = \gamma > 1$

2.10

Specifically, in classical statistical thermodynamics it is shown that

$$\gamma = \frac{N+2}{N}$$

2.11

Where N is the number of degrees of freedom of a molecule. For a classical monatomic gas, N=3 and

$$\gamma = \frac{5}{3} = 1.67$$

For a diatomic gas such as oxygen or nitrogen, N=5. (three displacement, two angles)

$$\gamma = \frac{7}{5} = 1.4$$

These results do not account for certain quantum mechanical effects. For example, in a diatomic molecule, there is a vibrational degree of freedom which is not of significance at normal temperatures, but which, at elevated temperatures can contribute the effect of two additional degrees of freedom, for an effective γ of. $\frac{9}{7} = 1.286$

At high temperatures other effects such as ionization and dissociation cause sufficient deviation from perfect gas behavior that it becomes difficult to define a unique value of γ .

In the absence of losses the energy equation can be found by noting that, in a steady state, the net energy flow into a given volume is zero. In our one dimensional formulation, this becomes:

$$\frac{d}{dx} \left(U + \frac{1}{2} u^2 + \frac{P}{\rho} \right) = \frac{d}{dx} \left(h + \frac{1}{2} u^2 \right) = 0 \quad 2.12$$

The quantity $h + \frac{1}{2} u^2$ is called the total enthalpy of the flow.

We will examine equation (2.12) for the case of a perfect gas. Equation (2.12) then becomes:

$$\frac{1}{2} u^2 + C_p T = C_p T_0 = \frac{1}{2} u_0^2 \quad 2.13$$

Here T_0 is called the total or stagnation temperature and is the value achieved when $u=0$. u_0 is called the limiting speed and is the speed which would be reached if $T=0$. u_0 is actually an upper bound for the effective exhaust velocity C which was defined in expansion through a convergent divergent nozzle from an effectively motionless combustion chamber. By examining this relation, it is rather easy to show some of the important parameters associated with a rocket fuel, since we might expect that specific impulse is almost directly proportional to u_0 , and high specific impulse would then be related to a high value of $C_p T_0$. The attainable temperature T_0 is obviously a function of the particular propellant and, but available adiabatic flame temperatures tend to run into a region of about 5000° F. This is obviously a high temperature for most available structural materials, and considerable effort is needed to assure proper thermal protection for the structure.

$$\text{now. } C_p = \frac{\gamma}{\gamma-1} R = \frac{\gamma}{\gamma-1} \frac{K}{m} \quad 2.14$$

K is the Boltzmann constant, and m is the molecular weight of the gas under consideration.

It is evident that high values of C_p correspond to low values of m , and to γ close to one.

Consider a calculation in which $m=25$, $T_0 = 5400^\circ \text{ R}$ and $r=1.21$.

$$2 C_p T_0 = 2 \times \frac{1.21}{1.21} \times 1990 \times 5400 = 1.24 \times 10^8 \frac{\text{Ft}^2}{\text{Sec}^2}$$

$$u_L = 11,100 \text{ Ft/Sec.}$$

If $C = u_L$.

$$\text{Isp} = \frac{11,100}{32.2} = 345 \text{ Sec.}$$

Let us consider the structure of a convergent divergent nozzle flow. In order to examine this, we consider the concept of Mach number and speed of sound.

The speed of sound in a gas is defined by:

$$a^2 = \frac{\partial p}{\partial \rho} \Big|_s \quad \text{and for a perfect gas,} \quad a^2 = \frac{\gamma p}{\rho} \quad \text{This equation}$$

for the propagation speed may be obtained from the acoustic approximation and its derivation is given in almost any book in fluid mechanics. Specifically Liepman and Roshko give such a derivation.

We define the Mach number to be the local ratio of speed u to acoustic speed a . Note that a is a function of position. If $M > 1$, the flow is supersonic, and $M < 1$ implies subsonic flow. The two regimes are considerably different in character, but this difference will not be discussed here.

The energy equation, may be written in the form-

$$\frac{1}{2} u^2 + C_p T = C_p T \left(1 + \frac{\gamma-1}{2} M^2 \right) = C_p T_0 \quad 2.15$$

So that $\frac{T_0}{T} = \left(1 + \frac{\gamma-1}{2} M^2 \right) \quad 2.16$

and

$$u = M a = M \sqrt{\gamma R T} = M \sqrt{\gamma R T_0} \left(1 + \frac{\gamma-1}{2} M^2 \right)^{-1/2}$$

If the flow is isentropic.

$$\frac{p_0}{p} = \left(1 + \frac{\gamma-1}{2} M^2 \right)^{\frac{\gamma}{\gamma-1}}$$

and

$$\frac{\rho_0}{\rho} = \left(1 + \frac{\gamma-1}{2} M^2 \right)^{\frac{\gamma}{\gamma-1}}$$

} 2.17

Where p_0 and ρ_0 correspond to the chamber conditions.

Take the logarithmic derivative of continuity equation (2.1) to obtain

$$\frac{d\rho}{\rho} + \frac{du}{u} + \frac{dA}{A} = 0 \quad 2.18$$

If the flow is isentropic

$$\frac{d\rho}{\rho} + \frac{du}{u} = \frac{(1-M^2) dM}{M(1+\frac{\gamma-1}{2} M^2)} \quad 2.19$$

And we note that, at the throat of a convergent divergent nozzle either $M=0$ or $M=1$. These two cases correspond to the situation where the Mach number achieves a maximum at the throat ($dM=0$) or $M=1$ at the throat, and no constraint is placed on dM at that point.

Note that if $M > 1$, dM and dA have the same sign, while if $M < 1$, dM and dA have opposite signs. If the flow starts from $M=0$, we then have two possibilities:

1. It never becomes supersonic, but rather stays subsonic all the way with the maximum Mach number occurring at the throat.
2. The flow remains subsonic until the throat is reached, and will become supersonic after the throat. This latter case is the supersonic nozzle representation.

One distinguishes between the two cases on the basis of the ambient pressure into which the gas exhausts.

Let us calculate. Suppose $\gamma=1.4$ and the flow exits at Mach 5. What is the pressure at the exit section?

$$P = P_0 \left(1 + \frac{\gamma-1}{2} M^2\right)^{-\frac{\gamma}{\gamma-1}} = P_0 (1 + 2 \times 25)^{-3.5} = .0045 P_0$$

For this Mach number to be achieved isentropically the ambient pressure must be as low as the exit pressure. If the ambient pressure is below P_{exit} , the nozzle is said to be underexpanded. If the ambient pressure is above this value it is overexpanded.

If the flow is to remain isentropic, the area ratio can be determined as a function of Mach number. Let A^* designate the throat area. Then:

$$\frac{A}{A^*} = \frac{\rho^* u^*}{\rho u} = \frac{(\rho^*/\rho_0) (T^*/T_0)^{1/2}}{(\rho/\rho_0) M (T/T_0)^{1/2}} = \frac{1}{M} \left(\frac{2}{\gamma+1} \left\{ 1 + \frac{\gamma-1}{2} M^2 \right\} \right)^{\frac{\gamma+1}{2(\gamma-1)}} \quad 2.20$$

$$\text{at } M=4 \quad \frac{A}{A^*} = 10.7 \quad \text{for } \gamma=1.4$$

Now consider the thrust on an engine, in a vacuum;



$$\text{Thrust} = \int (P + \rho u^2) dA \quad 2.21$$

exit area

now

$$\begin{aligned} \rho &= \rho_0 \left(1 + \frac{n-1}{2} M^2\right)^{-\frac{1}{\gamma-1}}, & u &= Ma = M \sqrt{\gamma R T_0} \left(1 + \frac{n-1}{2} M^2\right)^{-\frac{1}{2}} \\ P &= \rho_0 \left(1 + \frac{n-1}{2} M^2\right)^{-\frac{\gamma}{\gamma-1}}, & \rho u^2 &= \gamma M^2 P \\ (P + \rho u^2) A &= P(1 + \gamma M^2) A. \end{aligned} \quad 2.22$$

With the quasi-one dimensional analysis, in the presence of an ambient pressure

P_a , the thrust is represented as

$$T = \{P(1 + \gamma M^2) - P_a\} A \quad 2.23$$

Where $P_a A$ represents the ambient base pressure which is not recovered in the exit plane.

How should we operate with a given mass flow, and what are the trade offs?

If the burning chamber temperature is fixed, say T_0 then $P_0 = \rho_0 R T_0$. The mass flow rate can be defined by the throat mass flow.

$$m = \rho^* A^* U^* = \rho_0 \left(\frac{n+1}{2}\right)^{-\frac{1}{\gamma-1}} A^* \sqrt{\gamma R T_0} \left(\frac{n+1}{2}\right)^{-\frac{1}{2}} \quad 2.24$$

For a given mass flow and given T_0 , $\rho_0 A^*$ is fixed and $P_0 A^*$ is fixed.

For the same mass flow rate and T_0 , two similar jets (same area ratio) would deliver the same thrust (same $P_0 A^*$ is fixed) in a vacuum

However, in the presence of an ambient pressure, the loss ($P_a A$) is proportional to area. This would imply a high pressure small area system would be ideal. There are obvious structural limitations to this approach.

Suppose that the throat area A^* is given. Since the ambient loss is proportional to area, one would expect that an optimum expansion ratio would exist. In the quasi-one dimensional perfect gas framework, the problem is quite direct. If the thrust is expressed as a function of exit Mach number M_e , throat area A^* , chamber pressure P_0 , and ambient pressure P_a , then $\frac{dT}{dM_e}$ may be set equal to zero. There results the fact that for maximum thrust, $\frac{P_0}{P_a} = \left(1 + \frac{n-1}{2} M_e^2\right)^{\frac{n}{n-1}}$. Consequently, the optimum expansion ratio is that in which the exit plane pressure equals the ambient pressure.

In practice, the quasi-one dimensional perfect gas analysis is optimistic in its prediction of thrust. Various losses occur. For example, the exit flow will not be parallel, and thrust is lost through misdirection. The flow exerts a shear stress on the nozzle such that thrust is lost. The gas itself is not perfect and there will be losses due to non equilibrium phenomena. In particular, with two phase flow, there will be losses due to thermal and momentum lags.

Energy Generation in Rockets

By far, the most common form for a rocket engine is the chemically driven rocket. These rockets are generally classed in two categories, solid and liquid. In the solid rocket, the fuel and oxidizer are cast together in a nearly solid rubberlike form. The liquid rocket system is usually a bipropellant system in which the fuel is stored in one tank and the oxidizer in a second tank. The fuel and oxidizer are pumped, mixed, and injected into a combustion chamber where they are burned. There are advantages to each system, and each has its area of application.

The liquid engine has to date, achieved higher specific impulses, and is easier to control in the sense that thrust levels may be varied by control of fuel flow rates, and shut off and restart are commonly achieved.

The solid engine has distinct advantages in simplicity (lack of plumbing) and general storeability, but it has disadvantages in that the thrust level cannot be easily controlled, that the thrust level is often strongly influenced by temperature, that cut off is at least more costly and that restart is virtually impossible.

The physical characteristics of many propellants are listed in Glass tone, and will not be repeated here.

The primary objective is to achieve high specific impulse, or high combustion temperatures with low molecular weight. Hydrogen fuel and either oxygen or flourine oxidizer yields specific impulses of the order of 400 sec.

Solid propellants as yet do not deliver specific impulse at this level. A figure of 300 sec. would likely be very good. Many of the high Isp rockets are loaded with metal which burns. The products of combustion then include particles of metallic oxide. The presence of these particles modifies nozzle design, so as to minimize thermal and momentum drag.

A solid propellant normally requires a given pressure level to sustain burning. Ignition is effected by firing an igniter which sets up the proper pressure temperature environment. The burning rate ($16/\text{Sec Ft}^2$) can be correlated with pressure such that the rate is proportional to P^k , $k \approx 0.7$. A high pressure environment then induces more rapid burning. In order to sustain a given thrust, the rocket grain should be designed that the exposed area is roughly invariant with time. This, combined with heat transfer considerations is the reason for the star shaped cutout in some grain designs.

Various exotic schemes have been proposed. One, using a nuclear reactor, passes a nonreacting gas through a heat exchanger, which is heated by the nuclear reactor. A low molecular weight gas can be used and, at temperatures equivalent to those of chemical reactions, much higher specific impulses can be induced (velocity $\sim \sqrt{\frac{1}{m}}$). The weight of the reactor and shielding can be quite costly.

Ion engines are devices which accelerate ions electrostatically. Very high exit speeds can be induced, but so far, only very low thrust levels have been achieved. There is a significant problem in space charge neutralization.

Operation of Rocket Systems

We will be concerned with the question of optimum burning of fuel in order to achieve some objective. First, consider the problem of optimum thrusting of a vertically climbing rocket, with no aerodynamic drag, a constant specific impulse, and a constant gravitational field (flat earth).

$$m\dot{v} = -\dot{m}C - mg. \quad 2.25$$

$$\text{or } \dot{v} = -\frac{\dot{m}}{m}C - g \quad 2.26$$

$$\text{hence: } v = \frac{dy}{dt} \quad (y \text{ is altitude})$$

$$\frac{dv}{dt} = \frac{dv}{dy} \frac{dy}{dt} = v \frac{dv}{dy}$$

Then equation(2.25) can be written in the form

$$\frac{dv}{dy} \log me^{\frac{v}{C}} = -\frac{g}{C} \quad 2.27$$

The quantity $\log me^{\frac{v}{C}}$ can be considered to be a measure of how fast a vehicle could go if its remaining fuel were expended instantaneously (so that gravity impulse would have no effect). With such a program, $\log me^{\frac{v}{C}}$, a $\log m + \frac{v}{C}$ would remain constant, and the maximum achievable speed would be given by $v_{max} = v + C \log \frac{m}{m_{min}}$.

The optimum thrust program may be shown to be that which maximizes $\log me^{\frac{v}{C}}$ at every altitude since, the optimum could be exceeded by an impulsive burn otherwise. Obviously in this case, the optimum procedure is one of firing as rapidly as possible. The idea would be to fire instantaneously.

One can modify the program to account for aerodynamic drag by noting that, with drag:

$$m\dot{v} = -\dot{m}C - mg - \frac{1}{2}\rho v^2 S C_D \quad 2.28$$

(here ρ is a function of altitude)

Then, by similar manipulation,

$$\frac{d}{dt} (\log m e^{\xi}) = -\frac{g}{c v} - \frac{1}{2c} \frac{\rho v S C_D}{m} \quad 2.29$$

Again, the optimum procedure minimizes the loss in $m e^{\xi}$ at each station, and results in a velocity program such that, $\frac{\partial}{\partial v} (-\frac{g}{c v} - \frac{1}{2c} \rho v S C_D) = 0$.

For a constant C_D , one obtains: $\frac{v}{c} \ll 1$
 $\frac{g}{v^2} = \frac{1}{2m} \rho S C_D \quad 2.30$

This is a speed such that the drag equals the weight of the vehicle. Such a program will call for an impulsive burn until the drag equals the weight, followed by an acceleration such that the decrease in density is counteracted by the increase in speed such that the drag-weight balance is preserved. One must be careful to check that \dot{m} is negative. (Note- The variable $m e^{\xi}$ has been used by several authors, but, to the best of my knowledge, it was first used by H. Lewey in 1943. R.HE).

Next consider a simplified problem of a flat earth where a vehicle is boosted by a constant acceleration thrust. It is desired to place the vehicle into horizontal flight with speed v_1 , at an altitude h_1 . What is the shortest time procedure for effecting this maneuver?

$$\left. \begin{aligned} \ddot{x} &= a \cos \theta \\ \ddot{y} &= a \sin \theta - g \\ \text{at } t=0, \quad x=\dot{x}=y=\dot{y} &= 0 \\ \text{at } t=T, \quad y=h_1, \dot{x}=v_1, \dot{y} &= 0 \end{aligned} \right\} 2.31$$

As formulated, this is a problem in optimum control theory. When x, \dot{x}, y, \dot{y} are state variables and θ is a control variable, the problem is solved very simply by any of a variety of methods. The minimum value may be shown to be one wherein $\tan \theta$ is a linear function of time (Fried. Space Technology Lectures). This may be used as an indication of the proper programming of thrust angle as a function of time for injection purposes.

50

N67. 8.0.4 6.3

SPACE SCIENCE - SPACE SYSTEMS ENGINEERING'

II

by

E. Kane Casani

PREFACE

These notes are to serve as background material for this course in System Engineering. While the course will be mostly oriented toward System Engineering as used at the Jet Propulsion Laboratory in accomplishing its space flight missions, material is presented to give the student a better background and insight into the type or classes of problems that can be encountered in the field.

The course will be basically divided into two parts. The first part will include an introduction to system engineering and develop some of the classic tools of system engineering, such as linear programming, game theory, probability theory, decision theory and logic.

The second part of the course will be devoted to some of the practical considerations of system engineering as used at the Jet Propulsion Laboratory. This will include in particular the design of a hypothetical 1971 Mars Mission.

E. Kane Casani

TABLE OF CONTENTS

System Engineering

Linear Programming

Decision Theory

Astrodynamics

Planetary Approach

System Engineering

In the past two decades the field of System Engineering has grown immensely. The initial catalyst which has given this field its start was World War II. This is where operations research found its real beginning and tools like linear programming found a practical application. The war represented a rather large and complex operating system and was so difficult to understand that decisions could not be made only on intuition, but required some rather rigorous analysis. It was these types of problems which brought forth the operations research.

One of the by-products of the war is the missile and out of that has grown the Aerospace Industry as we know it today. This industry has been most instrumental in the further development of system engineering. Today the term system engineering is widely used and accepted throughout industry and there are now several Colleges and Universities offering courses and degrees in it.

Although the term is widely used and accepted, it is often misused and misunderstood. The terms system engineering, system analysis, operations research and operation analysis are often confused and used interchangeably. Let us look at these four terms and define them as we will use them throughout this course. These may not be the same definitions used by other people, but they are what we will use and therefore in our world they are correct. Since our world is the only correct world these then are the only acceptable definitions.

The term system engineering is used as a catch-all but should be considered in one of the following domains:

- 1) Operations Research
- 2) Operations Analysis
- 3) System Engineering
- 4) System Analysis

1) Operations Research

Operations Research is the development of the mathematical tools, such as linear programming, game theory, queueing theory, decision theory, etc. which are used throughout the field. The development of these tools is usually done on a higher level or at the research level. Their application is usually more classical than practical.

2) Operations Analysis

Operations Analysis is conducted within many branches of the military and large corporations such as the airlines, the oil industry, and the shipping industry to improve the operating efficiency of the organization. Mathematical models of an operating system are developed and analyzed. Some of the classic tools are used.

3) System Engineering

System Engineering is the conception, design, development, and operation of large complex systems. It deals with the design of an optimum system to perform a specific job within a set of specific constraints, the most important are usually performance, time, cost and manpower. It is this concept of System Engineering which is most familiar throughout the Aerospace Industry.

4) System Analysis

System Analysis deals with the analytical development models which describe the performance of various elements of a system. Tradeoff studies are an important part of this area. The development of the required software of the System is part of System Analysis.

When we use the word "System" it is necessary that we have a good understanding of exactly what that encompasses.

A system in the sense we are using it can be envisioned as shown in Fig. 1. Here we have an operating system with a set of definite performance requirements

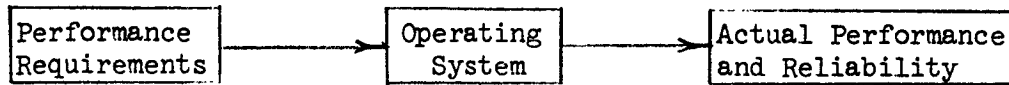


Figure 1

as its input and the actual performance and reliability as the output. In many cases the actual performance differs widely from the performance requirements, and the actual reliability is so low that system effectiveness makes it almost useless. In talking of these types of systems the word optimization is often used. The use of this term unbound is very misleading and confusing. There are two types of optimizing which should be kept in mind.

- 1) Fixed performance - minimum cost
- 2) Fixed cost - maximum performance

Here performance includes reliability. In the first case we fix the performance and then the optimum solution is one which meets the performance at the minimum cost. In the second case we fix the cost and then the optimum solution is one which maximizes the performance.

In referring to an operating system it is important to recognize all its major elements. The system consists of:

- 1) Hardware
- 2) Operating Procedures
- 3) Personnel and Training
- 4) Environment
- 5) Support Equipment and Logistics
- 6) Cost

System Engineering Process

Let us consider the process by which system engineering is actually conducted. System engineering is as defined earlier, and includes the conception, design, development and operation of a system. This process consists of the following

major steps:

- 1) Project Definition
- 2) System Design
- 3) Preliminary Design
- 4) Hard Design
- 5) Fabrication
- 6) Assembly and Test
- 7) Operation

Project Definition

There is some work which is actually conducted before the project definition which demonstrates the general requirement for this project and the general feasibility of the project. The method used in this work is usually rather unorthodox and the logic is often difficult to understand. There have been many projects proposed and undertaken which were ostensibly to fulfill some specific military or civic need, which were actually only to fill someone's pocket or ballot box. Be that as it may the problem we want to look at is given a set of project requirements, how does one go about accomplishing the project. These requirements are often stated in very general terms; "Take close up pictures of Mars in 1964" or "Put a man on the moon before 1970". It is often the case that the bigger the assignment the more general the requirements.

During the project definition phase the bounds and the groundrules under which the project is going to operate are defined. These include such things as the management tools, the overall schedule, the manpower and the budget. A set of agreed upon project objectives must be developed. These objectives should be clearly stated and understood by the project management and customer. This is a most important step, since it is against these stated objectives that the project

will be continually reviewed and finally appraised. The length of the project definition phase is usually extremely short in comparison to the life of the project, yet some of the most critical decisions which shape the entire project must be made during this phase.

System Design

The System Design Phase consists of defining a system capable of fulfilling the project objectives. This system should be defined in total with particular emphasis on those areas in which extensive development is anticipated. A general system level understanding of the entire system is required in which the level of understanding is generally the same throughout all elements of the system. This is not to imply that the detail mechanization of the subsystems are specified, but that the functional requirements of all subsystems is described. At the end of this phase there should not be any major areas uncovered, or any basic feasibility questions unanswered. If there are then it is foolhearted to proceed with the project, for when we are concerned with large complex systems they are very serious in nature and one missing major element is catastrophic on the final outcome. This activity is sometimes called the Conceptual Design Phase.

Preliminary Design

The Preliminary Design should always begin with a critical design review of the system design. After this review it may be necessary to modify the system design. During this phase the subsystems are functionally specified and their interface characteristics are specified. This is an extremely critical phase in the development of a system. To assure the best overall system design continual tradeoffs between the subsystems must be performed. A proper balance of risks must be achieved; this is best accomplished by prudent tradeoffs between the subsystems conducted at the system level. In addition, care must be taken to avoid strong intra-subsystem dependance. The system at this level of design

should have as much compliance as practical.

Hard Design

During this phase the detail design of each subsystem is carried out. These designs are developed in accordance with the functional specifications developed in the Preliminary Design. It is important that all subsystems enter this phase of the design at an equal level of design (i.e. good functional specifications). For if one subsystem is poorly understood, its design may be dictated by the others and this may present extreme difficulties in mechanization which could have been avoided by some proper tradeoffs earlier in the Preliminary Design.

Fabrication

This phase is the actual manufacturing and inspection of the hardware. The hardware dependent software is developed and checked out during this phase.

Assembly and Test

After fabrication the subsystems are tested at the subsystem level and then assembled into an entire system. System tests are then conducted to demonstrate the system's capability to meet its performance requirements. Specific tests are conducted to disclose any anomalies at the system level.

Operation

After the system is successfully working as a unit it is then committed to operations. Systems which have long operational lifetime and are a production line item, such as large missile systems, must have a well established feed back loop into the design as operational problems are disclosed.

Linear Programming

Linear Programming is a technique which has been developed since the end of World War II. It actually got its real start during the War. It is a technique which deals with certain types of problems. These problems are characterized in that they are linear in nature and usually have no one unique solution. In other words the problem can be bound by a set of linear constraints and within this set of constraints a specific function is to be optimized. Problems of this type also usually contain many variables.

The general linear programming problem can be reduced to the following formulation:

To find x_j (≥ 0), $j = 1, 2, 3, \dots, n$.

Subject to the following constraint

$$\sum_{j=1}^n a_{ij} x_j = b_i, \quad i = 1, 2, 3, \dots, m$$

where $m < n$

such that

$$C = \sum_{j=1}^n c_j x_j \text{ is a minimum (or maximum)}$$

and where a_{ij} , b_i and c_j are all given constants,

The equations which are of the form $\sum_{j=1}^n a_{ij} x_j = b_i$ are called

the constraint equations. The equation of the form $C = \sum_{j=1}^n c_j x_j$ is

called the cost function or sometimes the objective function. It is the value of this cost function which we wish to optimize.

The expanded form of the constraint equations is:

$$a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + a_{14} x_4 + \dots + a_{1n} x_n = b_1$$

$$a_{21} x_1 + a_{22} x_2 + a_{23} x_3 + a_{24} x_4 + \dots + a_{2n} x_n = b_2$$

$$a_{31} x_1 + a_{32} x_2 + a_{33} x_3 + a_{34} x_4 + \dots + a_{3n} x_n = b_3$$

$$\begin{array}{ccccccc} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

$$a_{m1} x_1 + a_{m2} x_2 + a_{m3} x_3 + a_{m4} x_4 + \dots + a_{mn} x_n = b_m$$

and the cost function:

$$c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4 + \dots + c_n x_n = C$$

It can be seen that if $m = n$ there are m equation and n unknowns. This is then a simple problem of m simultaneous linear equations, whose solution is straight forward. Now if $m > n$, that is there are more equations than unknowns, the problem is overconstrained and has no meaningful solution.

The case of interest is where $m < n$, that is there are more unknowns than equations. In such a situation we can arbitrarily chose any values for $n - m$ of the variables. This then reduces the problem to one which can be solved by the use of simultaneous equations. Then the trick is to chose these $n - m$ variables such that when the values of all the x_j 's are obtained. they give a minimum (or maximum) solution to the cost function.

There are techniques for solving three types of problems, possibly the most usually used is the Simplex Method. This method allows rigorous solutions to large problems of this nature, while smaller problems of this class can be solved by inspection and a few rules of thumb.

The Assignment Problem

Consider a large corporation with branch offices throughout the country. This corporation has four executive job openings, one in New York, Los Angeles, Philadelphia and Detroit. The corporation has been studying the records of its junior executives and has chosen four men to fill the positions. These men are located in branch offices scattered across the country, one in San Francisco, Chicago, New Orleans, and Milwaukee. Now all jobs are equal and all men are equally suited to fill any job. The problem is to determine the minimum cost of relocating the four men. Moving any one man from his present location to any one of the four job locations has associated with it a fixed cost. It is therefore possible to determine the cost of the sixteen possible moves. If we let the jobs be represented by J_j , $j = 1, 2, 3, 4$ and the men by M_i , $i = 1, 2, 3, 4$ then the problem can be represented in the matrix shown in figure 1. The cost numbers are shown in thousands of dollars.

	J_1	J_2	J_3	J_4
M_1	14	5	5	5
M_2	2	12	6	7
M_3	7	8	3	9
M_4	2	4	6	10

Figure 1

By inspection we can see that this problem is of the general form of a linear programming problem, with an additional consideration, we can only send one man to one job.

Now let c_{ij} be the cost index such that

c_{ij} is the cost of sending M_i to J_j , and let x_{ij} be the assignment index such that

$x_{ij} = 1$ if M_i is assigned to J_j and,

$x_{ij} = 0$ if M_i is not assigned to J_j

It must also be noted that

$$\sum_{j=1}^4 x_{ij} = 1, i = 1, 2, 3, 4$$

$$\sum_{i=1}^4 x_{ij} = 1, j = 1, 2, 3, 4$$

These two statements imply that one man can be assigned to one job only, that is x_{ij} can take the value one only once in each row and column.

Then the cost function has the form:

$$C = \sum_{i=1}^4 \sum_{j=1}^4 c_{ij} x_{ij}$$

and this double summation, subject to the previous constraints, is to be minimized. In this particular problem there are 16 unknowns, 7 dependent variables and 9 independent variables.

This problem can be solved by inspection and use of two general principles.

- 1) The Principle of Least Choice
- 2) The Principle of Interference.

Since the cost function is to be minimized the first obvious choice would be to pick the smallest cost element of the matrix. In column one

the c_{21} and c_{41} elements are both 2. As soon as one element is chosen, by the principle of interference, all other elements of its row and column are eliminated. If the c_{21} element is chosen then the remaining elements of the second row are eliminated. This appears to be a prudent choice because of the 12 in that row. The next smallest element of this matrix is c_{33} , the 3, and then c_{42} , the 4, and then c_{14} , the 5. Choosing each one of these elements yields the solution shown in figure 2.

	J_1	J_2	J_3	J_4
M_1	14	5	5	⑤
M_2	②	12	6	7
M_3	7	8	③	9
M_4	2	④	6	10

Figure 2

The value of the cost function for this solution is:

$$C = 2 + 4 + 3 + 5$$

$$C = 14$$

and this is the minimum. To prove that this solution is actually a minimum arrange all the c_{ij} 's in ascending order.

2, 2, 3, 4, 5, 5, 5, 6, 6, 7, 7, 8, 9, 10, 12, 14

The first c_{ij} is 2 but this is not independent of the second c_{ij} and therefore only one can be chosen. This dependence is because they are both in the first column. The next four c_{ij} 's are 2, 3, 4, 5 and their sum is 14 the value of C. Therefore, $C = 14$ must be the minimum solution.

At this point a short discussion on the principle of interference is proper. These types of problems are of a sequential decision class.

That is for each decision which is made there are a number of other choices (potential decisions) which are eliminated. It also turns out that the first decision eliminates the most choices, and each subsequent decision a smaller number. This is illustrated in the problem. The original matrix was a 4×4 matrix with 16 possible choices, when c_{21} was chosen the matrix was then reduced to a 3×3 matrix with only 9 possible choices. The number of available choices is reduced quadratically with each decision. A point to be noted here is the importance of the early decisions.

This method of solution can be used on many such transportation problems. As the size of the matrix increases the ease with which this method can be used reduces. Matrices larger than 10×10 are best solved by more formal techniques. One such technique is by the use of an established algorithm.

With several fundamental assumptions, which it is not within the scope of this text to prove, we will develop a technique for solving larger assignment matrices. First we will assume that if a matrix has a set (or sets) of independent elements whose sum is a minimum, that by adding or subtracting a constant from every element in any row (rows) or column (columns) we will generate a new matrix whose minimum is contained in the same original set of independent elements. A set of independent elements is one in which each element is contained in one and only row and column of the matrix.

Secondly, we will assume that by proper manipulation of the matrix we can develop a set of independent zeros in the matrix. Finally, this set of independent zero elements corresponds uniquely to a minimum solution of the original matrix.

Consider the matrix of the original problem, shown in figure 3.

	J_1	J_2	J_3	J_4
M_1	14	5	5	5
M_2	2	12	6	7
M_3	7	8	3	9
M_4	2	4	6	10

Figure 3

Now subtracting the smallest element from each column a new matrix is developed (figure 4). This matrix contains a set of independent zeros which do correspond to the original solution as shown in Figure 2.

	J_1	J_2	J_3	J_4
M_1	12	1	2	0*
M_2	0*	8	3	2
M_3	5	4	0*	4
M_4	0	0*	3	5

Figure 4

This serves to illustrate the general idea, but with a rather simple matrix. Figure 5 shows another matrix in which the answer will not fall out quite so easy.

2	6	5	9
3	4	8	8
5	1	2	3
4	3	2	7

Figure 5

First subtract from each row its smallest element (see Fig. 6).

0	4	3	7
0	1	5	5
4	0	1	2
2	1	0	7

Figure 6

This does not give zeros in all columns, so subtract from each column its smallest element (see Fig. 7).

0	4	3	5
0	1	5	3
4	0	1	0
2	1	0	5

Figure 7

There now exists at least one zero in each row and column, but only three of them are independent. Now to create another independent zero, add 1 to the third row (see Fig. 8) and then subtract

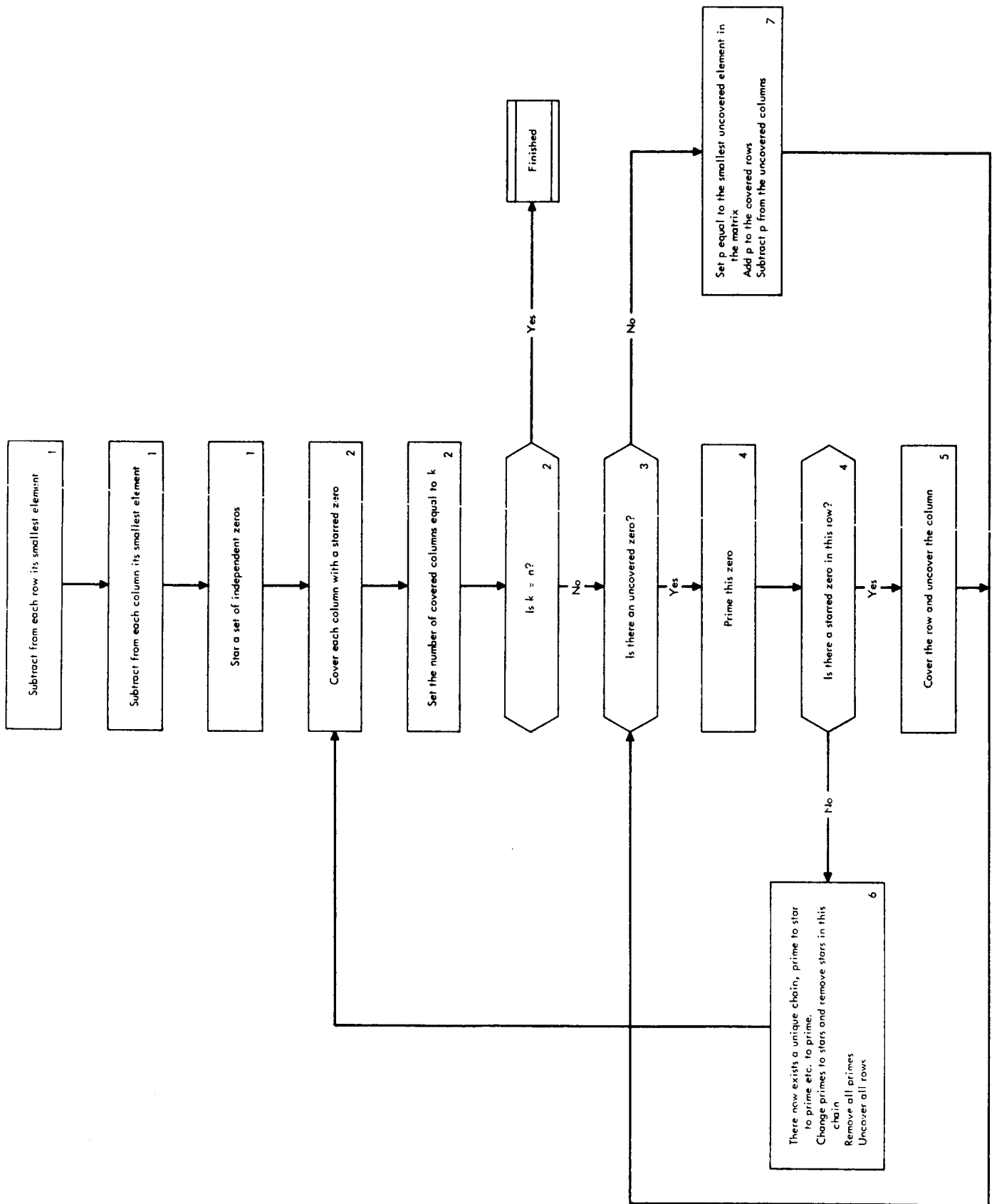
0	4	3	5
0	1	5	3
5	1	2	1
2	1	0	5

Figure 8

from each column its smallest element (see Fig. 9). This matrix now contains

0*	3	3	4
0	0*	5	2
5	0	2	0*
2	0	0*	4

Figure 9



ALGORITHM FOR ASSIGNMENT PROBLEM

Fig. 10

four independent zeros. These zeros now correspond to the elements of the original matrix whose sum is a minimum. The solution then is

$$C = 2 + 4 + 2 + 3$$

$$C = 11$$

Let us now look at the application of the algorithm shown in figure 10 to this problem. This algorithm can be stated in the following seven steps. When these steps are followed properly the solution will be achieved.

1. Subtract from each row its smallest element. Subtract from each column its smallest element. Choose a trial set of independent zeros and star them. Go to step 2.
2. Cover each column which has a starred zero. If all columns are covered, the solution is complete; if not, go to step 3.
3. Look for an uncovered zero. If there is none, go to step 7. If one is found, go to step 4.
4. Prime this zero. Look for a starred zero in the same row. If there is one, go to step 5. If there is none, go to step 6.
5. Cover the row, uncover the column of the starred zero, and go to step 3.
6. There now exists a unique chain, starting at the primed zero, going vertically to a starred zero, horizontally to a primed zero, etc., and ending on a primed zero (with no starred zero in its column). Go through this chain, changing primes to stars and erasing stars. Now erase all primes, uncover all rows, and go to step 2.
7. Find the smallest uncovered element in the matrix. Add this

element to the covered rows and subtract it from the uncovered columns (or add it to the covered columns and subtract it from the uncovered rows) (or add it to the twice-covered elements and subtract it from the uncovered elements). Do not change any stars, primes, or coverings. Go to step 3.

The original problem is shown in figure 11.

2	6	5	9
3	4	8	8
5	1	2	3
4	3	2	7

Figure 11

Subtract from each row its smallest element (figure 12)

0	4	3	7
0	1	5	5
4	0	1	2
2	1	0	7

Figure 12

Subtract from each column its smallest element (figure 13)

0	4	3	5
0	1	5	3
4	0	1	0
2	1	0	5

Figure 13

Chose a trial set of independent zeros and star them (fig. 14)

0*	4	3	5
0	1	5	3
4	0*	1	0
2	1	0*	3

Figure 14

Cover each column which has a starred zero(fig. 15)

x	x	x	
0*	4	3	5
0	1	5	3
4	0*	1	0
2	1	0*	3

Figure 15

Look for an uncovered zero. One exists in the fourth column. Prime this zero. A starred zero exists in the same row. Now cover the row and uncover the column of the starred zero (see fig. 16).

	X		X	
	0*	4	3	5
	0	1	5	3
X	4	0*	1	0'
	2	1	0*	3

Figure 16

Now all zeros are covered, step seven now requires us to find the smallest uncovered element of the matrix. This is 1. Add this element to the covered rows (see fig. 17), and

	X		X	
	0*	4	3	5
	0	1	5	3
X	5	1*	2	1'
	2	1	0*	3

Figure 17

subtract it from the uncovered column (see fig. 18).

	X		X	
	0*	3	3	4
	0	0	5	2
X	5	0*	2	0'
	2	0	0*	2

Figure 18

Look for an uncovered zero. One exists in the second column, second row. Prime this zero (fig. 19)

	X		X	
	0*	3	3	4
	0	0'	5	2
X	5	0*	2	0'
	2	0	0*	2

Figure 19

There are no starred zeros in the second row. There now exists a unique chain, starting at the primed zero, going vertically to a starred zero, horizontally to a prime zero. Go through this chain, changing primes to stars and removing stars (see fig. 20)

0*	3	3	4
0	0*	5	2
5	0	2	0*
2	0	0*	2

Figure 20

This then (fig. 20) is the solution to the problem since there are four independent zeros. The minimum solution from the original matrix is then

$$C = 2 + 4 + 2 + 3$$

$$C = 11$$

The Transportation Problem

The transportation problem is another classic problem of linear programming. It is similar to the assignment problem in many ways. To get an insight into the problem let us consider the following situation.

A large bicycle manufacturer has three factories located in three different spots across the country, and at five other locations he has his warehouses. Now the cost of shipping one bike from any factory to any warehouse is unequally determined. The problem we wish to concern ourselves with is the minimum cost of shipping all the bikes from the factories to the warehouse. Where it is assumed that each warehouse has a specific demand, the number of bikes wanted, and each factory has a specific supply, the number of bikes available. To simplify the problem let us consider only the case where the total supply is exactly equal to the total demand.

Let us assign some specific values to the problem and develop a solution.

		r_j				
		W_1	W_2	W_3	W_4	W_5
a_i		40	40	80	80	80
F_1	100	2	5	2	3	3
F_2	100	2	2	2	1	0
F_3	120	3	6	2	1	4

Here the problem is represented in a matrix form.

The a_i 's are the amounts required; the r_j 's are the amounts required; the elements are the cost of shipping from each factory to each warehouse per unit. Then in general we have:

F_i is the i^{th} factory,

a_i is the amount available at the i^{th} factory,

W_j is the j^{th} warehouse,
 r_j is the amount required at the j^{th} factory,
 C_{ij} is the unit cost of shipping from F_i to W_j ,
 x_{ij} is the number of units shipped from F_i to W_j ,

$$\sum_{i=1}^3 a_i = \sum_{j=1}^5 r_j,$$

$$\sum_{j=1}^5 x_{ij} = a_i, \quad i = 1, 2, 3, \text{ and}$$

$$\sum_{i=1}^3 x_{ij} = r_j, \quad j = 1, 2, 3, 4, 5.$$

The problem is now to minimize the total cost of shipping all the units. That is to minimize the following:

$$C = \sum_{i=1}^3 \sum_{j=1}^5 C_{ij} x_{ij}.$$

Let us first find an initial trial solution which satisfies the constraints, and then inspect it for optimality. First find the lowest cost coefficient, in this case it is $c_{25} = 0$, then assign the maximum number of units, in this case 80. Now in the second row there are 20 more units available. Assign these to the lowest cost coefficients in that row, making sure to observe the column constraint. In this case we can assign all 20 to the fourth column. Then in this column look for the lowest cost coefficient and assign the remaining 60 of the fourth row. We proceed in this fashion through the matrix from row to column making assignments until all units are assigned. The assignment is as shown below:

		r_j	W_1	W_2	W_3	W_4	W_5
		a_i	40	40	80	80	80
F_1	100		40	40	20	0	0
F_2	100		0	0	0	20	80
F_3	120		0	0	60	60	0

Now let us test this solution for optimality by inspecting what would happen to this total if we reassigned one unit to a place where we now have a zero. For example let us see what would have happened if we made an entry at x_{35} . If we add one to x_{35} , we must subtract one from x_{25} , add one to x_{24} and finally subtract one from x_{34} . This shift would not violate the constraints and change the cost by

$$c_{35} - c_{25} + c_{24} - c_{34} = \Delta C$$

$$4 - 0 + 1 - 1 = +4.$$

Thus we see that this shift will increase the cost by 4 for every unit we ship from F_3 to W_5 . By using this method we can check the entire solution. If we find a shift which produces a negative total then we would transfer as many units as possible through that shift.

Since c_{12} is the highest cost coefficient let us look at shifting some units out of there. Let us try for a positive entry at x_{22} then the shift would alternately add and subtract to x_{22} , x_{24} , x_{34} , x_{33} , x_{13} and x_{12} or the cost would change by

$$c_{22} - c_{24} + c_{34} - c_{33} + c_{13} - c_{12}$$

$$2 - 1 + 1 - 2 + 2 - 5 = -3$$

This path is constrained by x_{24} (or x_{13}) which are 20, therefore we can transfer 20 units through this shift.

This new assignment is shown below:

rj \ a _i		W ₁	W ₂	W ₃	W ₄	W ₅
		40	40	80	80	80
F ₁	100	40	20	40	0	0
F ₂	100	0	20	0	0	80
F ₃	120	0	0	40	80	0

This procedure could be carried out for every unoccupied place, but could become rather lengthy. The method of shadow costs allows us to quickly inspect a solution and determine if it is the minimum or not, and if not what transfer to make. We will define shadow costs, u_i for every row and v_j for every column such that $u_i + v_j = c_{ij}$ for each non zero x_{ij} . Let us look at the first transfer we considered,

$$c_{35} - c_{25} + c_{24} - c_{34} = \Delta C \text{ or using}$$

the shadow costs for all but the zero x_{ij} we have

$$c_{35} - u_2 - v_5 + u_2 + u_4 - u_3 - u_4 = \Delta C$$

$$c_{35} - (v_5 + u_3) = C$$

Thus we see that if the shadow cost of the unoccupied cell does not exceed the true cost in that cell then the total cost will increase by occupying that cell. Conversely if the shadow exceeds the true cost then an improvement can be realized.

$$c_{ij} > u_i + v_j \quad \text{no shift}$$

$$c_{ij} < u_i + v_j \quad \text{shift}$$

In determining the values of the u_i 's and v_j 's we assign an arbitrary value to one of them (usually set $u_1 = 0$) and work our way through the matrix

determining all other values of u_i and v_j , remembering only to evaluate u_i and v_j for the occupied cells. This can be seen more clearly by looking at the initial assignment.

40	40	20	0	0
0	0	0	20	80
0	0	60	60	0

We can now write the c_{ij} 's for the occupied cells, and then evaluate the u_i 's and v_j 's by starting with $u_1 = 0$.

v_j	2	5	2	1	0
u_i					
0	2	5	2	x	x
0	x	x	x	1	0
0	x	x	2	1	x

With these values of u_i and v_j we can determine the shadow costs for the unoccupied cells.

v_j	2	5	2	1	0
u_i					
0	x	x	x	1	0
0	2	5	2	x	x
0	2	5	x	x	0

Here we see that $u_2 + v_2 > c_{22}$ and, therefore we should improve the solution by transferring through that cell.

40	20	40	0	0
0	20	0	0	80
0	0	40	80	0

Again we must inspect this solution for optimality. Now writing the c_{ij} 's for the occupied cells, and evaluating the u_i 's and v_j 's, starting with $u_1 = 0$.

v_j	2	5	2	1	3
u_i					
0	2	5	2	x	x
-3	x	2	x	x	0
0	x	x	2	1	x

And then the shadow costs for the unoccupied cells

v_j	2	5	2	1	3
u_i					
0	x	x	x	1	3
-3	-1	x	-1	-2	x
0	2	5	x	x	3

Now we see that all the shadow costs are less than the true costs and therefore the solution must be a minimum.

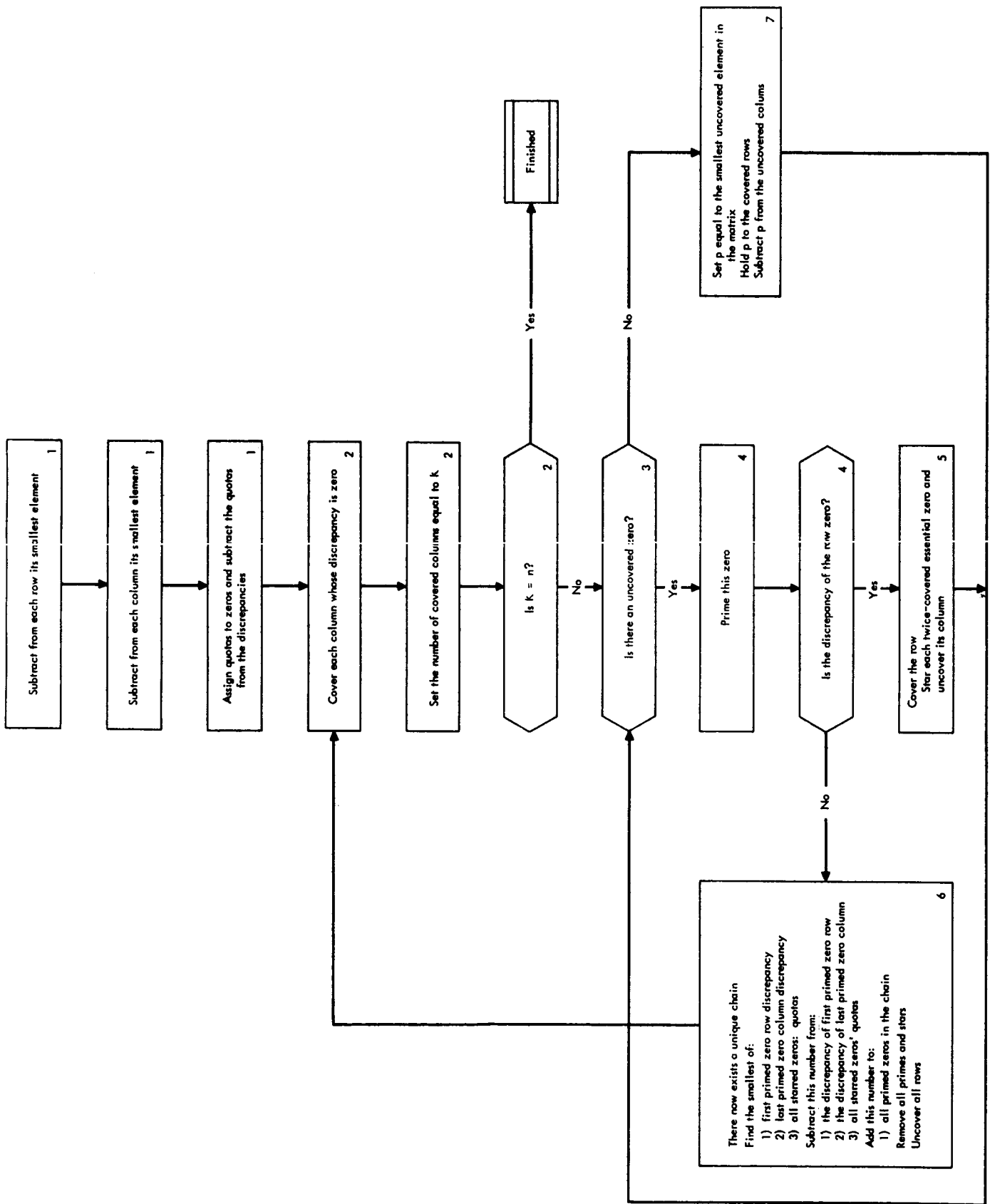
The transportation problem, like the assignment problem can also be solved by the use of an algorithm. The Algorithm is shown in Figure 1 and the steps required are:

Algorithm For Transportation Problem

1. Subtract from each row its smallest element. Subtract from each column its smallest element. Pick a trial set of quotas by assigning them to zeros, subtracting appropriately from the discrepancies. Go to step 2.
2. Cover each column whose discrepancy is zero. If all columns are covered, the solution is complete; if not, go to step 3.

3. Look for an uncovered zero. If there is none, go to step 7. If one is found, go to step 4.
4. Prime this zero. Check the discrepancy of the row; if it is not zero, go to step 6. If it is zero, go to step 5.
5. Cover the row, and for each twice-covered essential zero, star the zero and uncover its column. Go to step 3.
6. There now exists a unique chain, as above. Find the smallest of the following numbers: the discrepancy of the row of the first primed zero in the chain; the discrepancy of the column of the last primed zero in the chain; the quota of each starred zero in the chain. This number is to be subtracted from each of these two discrepancies, and from the quota of every starred zero in the chain, and to be added to the quota of every primed zero in the chain. Now erase all primes and stars, uncover all rows, and go to step 2.
7. Find the smallest uncovered element in the matrix. Add this element to the covered rows and subtract it from the uncovered columns (or add it to the covered columns and subtract it from the uncovered rows) (or add it to the twice-covered elements and subtract it from the uncovered elements.) Do not change any stars, primes, or coverings. Go to step 3.

Note: "Discrepancies" are amounts to be shipped which have not yet been assigned. "Quotas" are amounts which have already been assigned to particular elements of the matrix (i.e., particular routes.) An essential zero is one whose quota is greater than zero.



ALGORITHM FOR TRANSPORTATION PROBLEM

Fig. 1

It is interesting to notice that a transportation problem is in reality a special form of the assignment problem. This implies that if in a transportation problem a_i and r_j are integers all the x_{ij} 's must also be integers. To understand this observation let us define both the assignment problem and then the transportation problem in the general case.

Assignment Problem Definition

An assignment problem is of the form where k men are to be assigned to k jobs and the cost of assigning each man to each job is uniquely defined, and we wish to find the minimum assignment of all men. This problem can be represented in a matrix form as shown in Fig. 1.

	J_1	J_2	J_3	- -	J_j	- -	J_k
M_1	c_{11}	c_{12}	c_{13}	- -	c_{1j}	--	c_{1k}
M_2	c_{21}	c_{22}	c_{23}	- -	c_{2j}	- -	c_{2k}
M_3	c_{31}	c_{32}	c_{33}	- -	c_{3j}	- -	c_{3k}
M_i	c_{i1}	c_{i2}	c_{i3}	- -	c_{ij}	- -	c_{ik}
M_k	c_{k1}	c_{k2}	c_{k3}	- -	c_{kj}	- -	c_{kk}

Fig. 1

c_{ij} is the cost of assigning M_i to J_i

y_{ij} is the assignment index, such that

$y_{ij} = 1$ if M_i is assigned to J_j , and

$y_{ij} = 0$ if M_i is not assigned to J_j

$$\sum_{j=1}^k y_{ij} = 1, i = 1, 2, 3, \dots, k$$

$$\sum_{i=1}^k y_{ij} = 1, j = 1, 2, 3 \dots k$$

The cost function, C, to be minimized is

$$C_{\min} = \sum_{i=1}^k \sum_{j=1}^k c_{ij} y_{ij}$$

Transportation Problem Definition

A transportation problem is of the form where there are m destination and n sources, each source has a_i units available, and each destination has r_j units required. The cost of transporting from each source to each destination is uniquely defined, and we wish to find the minimum cost of transporting all units to the destinations. This problem can be represented in a matrix form as shown in Fig. 2.

$\begin{array}{c} r_j \\ \diagdown \\ a_i \end{array}$		D_1	D_2	D_3	---	D_j	---	D_m
		r_1	r_2	r_3		r_j		r_m
S_1	a_1	c_{11}	c_{12}	c_{13}		c_{1j}		c_{1m}
S_2	a_2	c_{21}	c_{22}	c_{23}		c_{2j}		c_{2m}
S_3	a_3	c_{31}	c_{32}	c_{33}		c_{3j}		c_{3m}
S_i	a_i	c_{i1}	c_{i2}	c_{i3}		c_{ij}		c_{im}
S_n	a_n	c_{n1}	c_{n2}	c_{n3}		c_{nj}		c_{nm}

Fig. 2

D_j is the jth destination

r_j is the number of units required at D_j

S_i is the ith source

a_i is the number of units available at S_i

c_{ij} is the cost of shipping one unit from S_i to D_j

x_{ij} is the number of units shipped from S_i to D_j

$$\text{If } \sum_{i=1}^n x_{ij} = r_j, j = 1, 2, 3 \dots m$$

$$\text{and if } \sum_{j=1}^m x_{ij} = a_i, i = 1, 2, 3 \dots n,$$

$$\text{then } \sum_{i=1}^n a_i = \sum_{j=1}^m r_j$$

This last statement implies the number of units required is exactly equal to the number available. While this may not always be true in the original problem, any problem can be formulated in this manner by the additional artificial sources or destinations.

The cost function, C, to be minimized is

$$C_{\min} = \sum_{i=1}^n \sum_{j=1}^m c_{ij} x_{ij}$$

Now with these definitions any transportation problem may be expanded into a large assignment problem as follows. Consider an m by n transportation problem where

$$\sum_{i=1}^n a_i = \sum_{j=1}^m r_j = k$$

This transportation problem can then be expanded into a k by k matrix which is composed of many subsets which are matrices with identical c_{ij} 's in each element and are a_i by r_j large ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$). This problem is now similar to a standard assignment problem and its optimum solution must contain values of $y_{ij} = 1$ or 0, where all the y_{ij} 's = 1 are independent. This large assignment problem can then be collapsed to the original transportation problem and the x_{ij} 's of the transportation will be the sum of the y_{ij} 's of the assignment problem for each constant c_{ij} subset of the matrix. Since the y_{ij} 's of the assignment problem can only have values 1 or 0, integer numbers, then

their sums, the x_{ij} 's of the transportation problem, must also be integers.

Decision Theory

The subject of decision theory has possibly more faces and interpretations than one can imagine. To gain some insight into the facade which we will concern ourselves, i.e. the more practical considerations, let us consider the following situation.

You are sitting on a park bench on a nice cool Sunday afternoon watching what ever seems to be the thing to watch. Then along comes a rather average young man, you hardly take notice of him, until he sits down next to you. Then you notice he has an ice cream cone in his hand. He offers you the ice cream cone. You have just finished eating and without too much thought reply "no thank you". You just don't feel like an ice cream cone. Now in refusing the cone did you make a decision? No not really, you simply reacted.

Now the stranger reaches into his pocket with his other hand and pulls out a small hand gun. He looks at you, places the gun in your side and says, "I think you want my ice cream cone". What do you do? You accept the ice cream cone rather quickly and thank him. Now, have you ~~made~~ made a decision? No, you had no other alternative but to accept. At least if you are a rational person, you had no other alternate. You could have chosen to get shot but that is hardly rational.

So there you are about to eat your ice cream cone, still not having made any decisions, when another man sits down along side of you. This second man whispers in your ear not to eat the ice cream cone. He claims that he knows the first man with the gun in your side and that this man goes around the park on Sunday afternoon passing out poisoned ice cream cones. He claims that this man is a little crazy.

Now you have a problem and now you have to make a decision. You also start asking yourself some questions which you must answer before you can make

any decision. Is the ice cream cone really poisoned? What kind of poison? Who is really crazy, the first man, the second man, both men or neither? Does the first man have a real gun? Is it really loaded? Will he really shoot?

Now what are your choices. You can eat the ice cream cone, hope nothing happens, get up, say good-bye and go to the nearest hospital to have your stomach pumped; you could eat it and take your chances; you could refuse to eat it and take your chances; or you could drop the ice cream cone and run for your life and hope the man is a poor shot.

It is this type of decision making which is to be discussed. Let us first define some of the major elements or key points of a decision, and then come back to this problem.

1) Problem

There must be a problem before we can talk about making a decision. There must be several choices at hand and the proper choice is not obvious. We must consider a decision as an irrevocable commitment of resources.

2) Uncertainty

To be concerned with a decision there must be uncertainty as to what the actual outcome will be. If the outcome is determined then no real decision is required.

3) Probability Theory

The fact that there is uncertainty involved implies that probability theory must play a strong part in decision theory.

4) State of Mind

While the theory of probability is underlying to entire approach to making a decision we must use the probability as our state of mind about the situation. The probability of an event occurring is a measure of our belief that it will occur.

5) Experience

The probability we associate with an event occurring is dependent on our experience.

6) Value

We must assign a value to each possible outcome of a situation. The value of an outcome must not be confused or influenced by the probability of the outcome occurring.

7) Risk Criteria

We must determine how much risk we are willing to take before we make the final decision. Do we want to maximize the expected value and reduce the probability of obtaining it, or do we want to minimize the probability of getting nothing?

8) Future

A decision must be influenced by the future. When a decision is made its outcome is dependent on things that happen subsequent to its being made.

9) Outcome

It is most important to realize that good decisions can have good outcomes or bad outcomes and that bad decisions can have good outcomes or bad outcomes.

It is interesting to note that Decision Theory is highly dependent on Probability Theory, yet Probability Theory is 350 years old while Decision Theory is only 20 years old.

Let us look at a problem in which a girl cannot decide where to have her wedding reception; inside, outside or on the porch. The real problem is rain or shine. Let's look at all possible outcomes and assign some value numbers to them.

		<u>Value</u>	<u>Expected Value</u>
inside	.6 rain	5	3.0
	.4 shine	4	1.6
porch	.6 rain	3	1.8
	.4 shine	6	2.4
outside	.6 rain	0	0.0
	.4 shine	10	4.0
			4.6
			4.2
			4.0

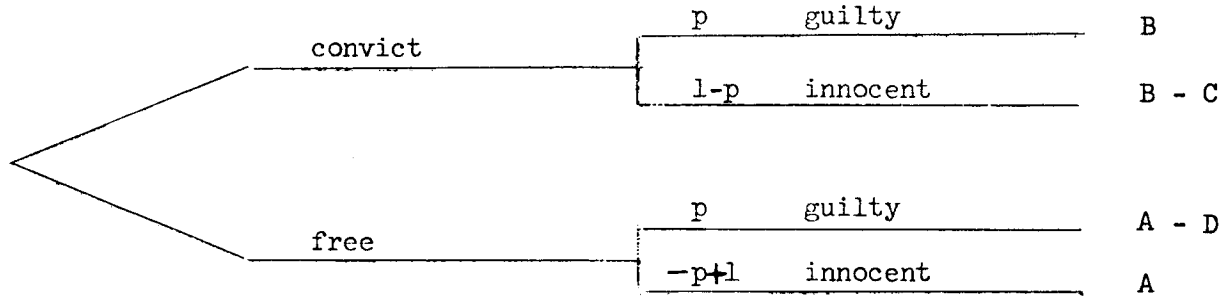
We can now determine the expected value of each decision (i.e. inside, outside, porch) which is the sum of the values times their probabilities.

	<u>Expected Value</u>	
<u>Probability of rain</u>	.4 to .6	.5 to .5
Inside	4.6	4.5
Porch	4.2	4.5
Outside	4.0	5.0

The above table shows how a small change in the probability of rain can change the expected value.

In addition to considering the Expected Value we must also consider our willingness to take a risk. With a 50 - 50 chance of rain our expected value is the highest if the party is outside, but 50% of the time we have a flop (i.e. value 0). Whereas, if the party is inside the expected value is 4.5 and the true value is never below 4.

Let us consider another problem in which a judge must decide whether to free or convict a man, given some probability, p , that the man is guilty and $1-p$ he is innocent.



A	Gain to society of freeing the man	\$7,000
B	Cost of keeping a man in jail	\$2,000
C	Cost of convicting an innocent man	\$100,000
D	Loss to society of freeing a guilty man	\$10,000

Now the Expected Value of each choice are

Convict $pB + (1-p) (B-C)$

Free $p (A-D) + (1-p) A$

Therefore the judge should convict if $pB + (1-p) (B-C) > p (A-D) + (1-p) A$

$$pB + B - C - pB + pC > pA - pD + A - pA$$

$$pC + pD > B + C + A$$

$$p > \frac{A+B+C}{C+D}$$

In the case shown p must be

$$p > \frac{7,000 + 2,000 + 100,000}{100,000 + 10,000}$$

$$p > \frac{109}{110}$$

$$p > .9910$$

Properties of Lotteries

1) Orderability

$A < B$ or $A > B$

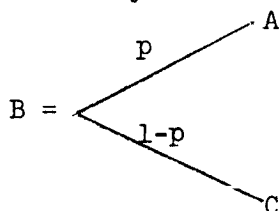
2) Continuity

If $A > B > C$ then there is some p such that

$$B \sim [p, A; (1-p) C]$$

$B \sim \tilde{B}$, is the certainty equivalent

This is to say that



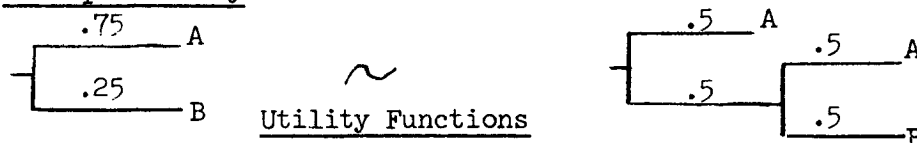
3) Substitutability

If $A = B$ then either can be chosen

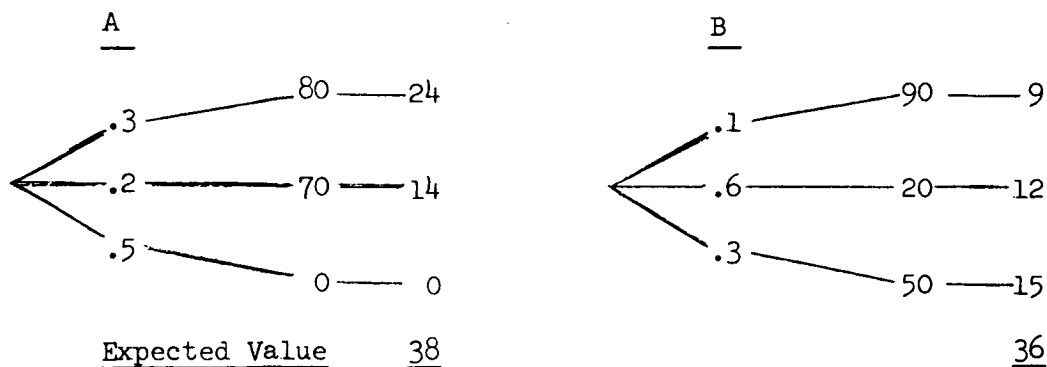
4) Monotonicity

If $A > B$, then chose A

5) Decomposability



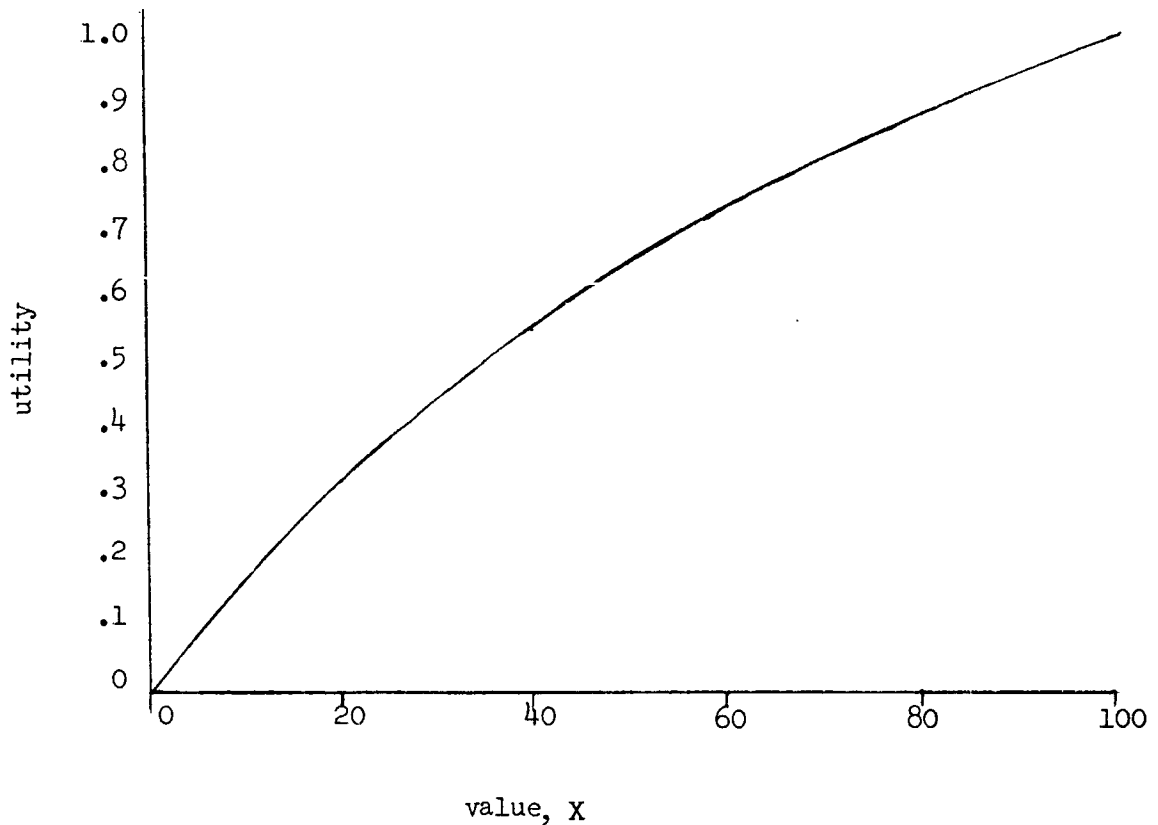
Let us look at two lotteries:



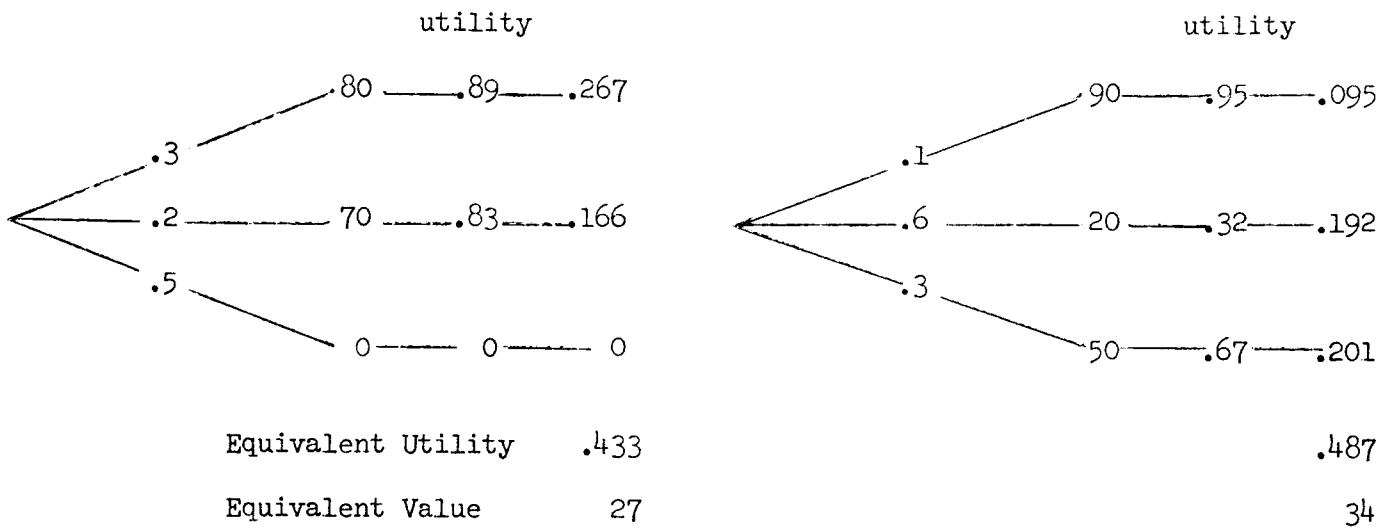
From the Expected Value only lottery A is the proper choice. It should be noted that in lottery A, 50% of the time I get nothing, while in lottery B 40% of the time I get 50 or better.

If I now consider the utility of these two lotteries, I must determine my risk character. Let us assume I am a risk averter with a utility function;

$$u(x) = 4/3 \left[1 - \left(\frac{1}{2} \right)^{\frac{x}{50}} \right]$$



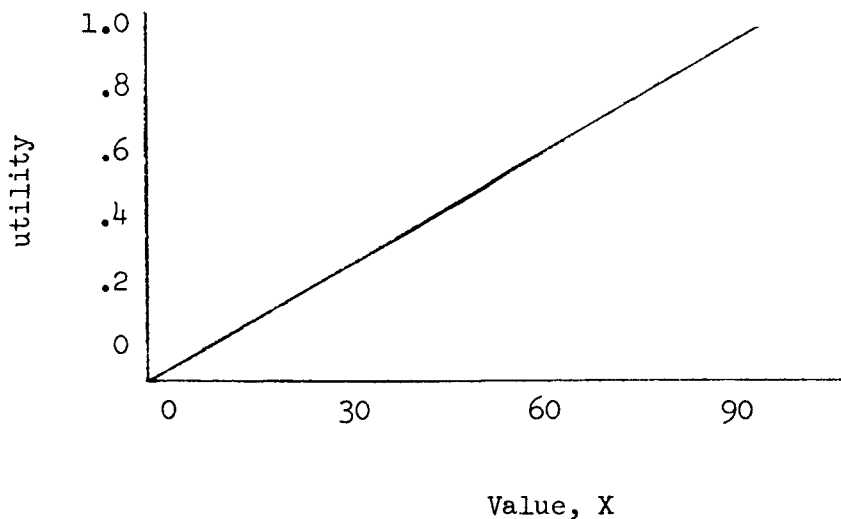
We can now determine the utility of both lotteries and then their value (for this risk averter).



We now see that the Equivalent Value is higher for lottery B than lottery A, but both are lower than the Expected Value. In other words for lottery A, with an expected value of 38, I would pay 27, and for lottery B, with an expected value of 36, I would pay 34.

Risk Indifferent

A risk indifferent person has a utility function with a one to one relationship between utility and value



slope = 1

Astrodynamics

With the advent of the space age the field of astrodynamics has grown rapidly. Most of the tools which are used today were essentially developed by Newton and others out of the basic discoveries of men like Kepler, Brahe and Copernicus. This work of Newton, later embellished by such mathematicians as Lagrange and Euler, composes what is classically called Celestial Mechanics. Astrodynamics, as we shall briefly study it, is engineering or practical application of Celestial Mechanics to the contemporary problems of space vehicles, exclusive of conventional aerodynamics and booster propulsion theory.

In dealing with the trajectories of an artificial satellite or an interplanetary spacecraft it is convenient to consider the general problem as a set of problems each of which can be considered as a two body problem. In the case of a trajectory from the Earth to a target planet (e.g. Mars), this is done by first considering the Earth as the center of the coordinate system, then the Sun, then Mars. By using this approach we can simplify what could be an extremely difficult problem. This simplification lets us get a close answer to the real problem and allows us to get a quick understanding of the situation.

Let us look first at the two body problem in its general form, then at the coordinate systems of interest and finally at some specific problems.

In astrodynamics when we talk about the two body problem we are restricting our thinking to motion of one (relatively small) body about another (relatively large) body. The second body is usually considered as the central force field, and is used as the center of the coordinate system in which the motion of the smaller body is described. For example, one can consider the motion of the Earth about the Sun. In this "two body problem" the Sun is the center force field, and the origin of an orthogonal coordinate system in which the Earth's motion can be described. In this system a surprisingly accurate

description of the Earth's motion can be developed without considering the effects, or perturbations of the other planets on the Earth.

There exists a fundamentally important relationship which uniquely describes the motion of a body in orbit about another body, called the vis-viva integral. This integral is commonly seen in the following form:

$$C_3 = V^2 - \frac{2GM}{R}$$

where: $GM = 3.9 \times 10^5 \frac{\text{km}^3}{\text{sec}^2}$ for Earth

R = the radius to the vehicle from the center of the Earth

V = velocity of the vehicle at a distance R

C_3 = twice the total geocentric energy per unit mass in km^2/sec^2 . C_3 is also the square of the hyperbolic excess velocity.

Another form of the vis-viva integral which is extremely useful and simple is the following dimensionless form.

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right)$$

All quantities are used in a dimensionless form, more will be said about this later.

r is the distance of the vehicle from the center of the system

\dot{s} is the speed of the vehicle at a distance r

a is the semi-major axis of the conic of the vehicle

It can be seen by inspection of this equation that it is the sum of the total potential and kinetic energy of the system. Specifically;

\dot{s}^2 corresponds to the kinetic energy

$\frac{2\mu}{r}$ corresponds to the potential energy

$-\frac{\mu}{a}$ corresponds to the total energy and is constant.

The study and fundamental understanding of this vis-viva integral is a most powerful tool. Its use in the conceptual design or system engineering of space ventures is almost unlimited. We will concern ourselves mainly with the use of this equation and forego its formal development.

Kepler's first law states, "The orbit of each planet is an ellipse with the sun at a focus." Newton later expanded this law to state that in all two body problems the motion under a central force field results in conic sections. The conic sections and some of their important constants and forms of the vis-viva integral are:

Circle

$$e = 0$$

$$a = r$$

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right)$$

$$\dot{s}^2 = \frac{1}{r}$$

Ellipse

$$0 < e < 1$$

$$a > 0$$

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right)$$

Parabola

$$e = 1$$

$$a = \infty$$

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right)$$

$$\dot{s}^2 = \mu \left(\frac{2}{r} \right)$$

Hyperbola

$$e > 1$$

$$a < 0$$

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{-a} \right)$$

$$\dot{s}^2 = \mu \left(\frac{2}{r} + \frac{1}{a} \right)$$

In astrodynamics and astronomy it is often useful, and more accurate, to use dimensionless system. This is the numbers which are actually used in the computation have no dimensions and are actually ratios to well known parameter. If we consider the quantities of length, speed, and mass we have the following basics to use;

Length

Geocentric

In considering systems in which the Earth is the central force field all linear dimensions are expressed in terms of the Earth's radius. Then in this system the distance to the surface of the Earth is;

$$r_{\oplus} = 1 = \frac{3957 \text{ mi}}{3957 \text{ mi}}$$

where we will assume the radius of the Earth is 3957 miles. The distance to the moon is, $r_{\lrcorner} = 60.3706$.

Heliocentric

In considering systems in which the Sun is the central force field all linear dimensions are expressed in terms of the semi-major axis of the Earth's orbit about the Sun. This distance is one astronomical unit (1 a.u.). This is approximately 92.90×10^6 mi. In this system the distance to all the planets is:

Mercury	0.3871
Venus	0.7233
Earth	1.0000
Mars	1.5237
Jupiter	5.2028
Saturn	9.5388
Uranus	19.1820
Neptune	30.0577
Pluto	39.5177

Speed

The speed is in terms of the satellite speed at a unit distance.

Geocentric

In this system the speed is the satellite speed at one Earth radius. The actual speed is 7.905 km/sec, 26,000 ft/sec, 4.912 mi/sec (≈ 5 mi/sec).

Heliocentric

In this system the speed is the satellite speed at 1 a.u., or the speed of the Earth in its own orbit. The actual speed is 29.6 km/sec, 96,700 ft/sec, 18.6 mi/sec.

Mass

The mass is expressed in terms of the most massive body in the system (i.e. the central body). In the vis-viva integral

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right)$$

is the sum of the two masses in the system in dimensionless form, $m_1 = 1$ and usually $m_1 > m_2$, therefore

$$\mu = m_1 + m_2 \simeq 1$$

Some useful Earth mass ratios are

Sun	331,950
Moon	0.012
Mercury	0.05
Venus	0.81
Earth	1.00
Mars	0.11
Jupiter	318.4
Saturn	95.3
Uranus	14.5
Neptune	17.2

Let us use these concepts and determine the altitude and speed of a 24 hour synchronous satellite in a circular orbit.

First we can determine the semi-major axis by the use of Kepler's third law.

$$\left(\frac{P_{\Delta}}{P_{\oplus}} \right)^2 = \left(\frac{a_{\Delta}}{a_{\oplus}} \right)^3$$

Where the sub Δ refers to the satellite and the sub \oplus refers to the Earth. Now since $a_{\oplus} = 1$ we can rewrite this expression

$$a_{\Delta} = \frac{P_{\Delta}^{2/3}}{P_{\oplus}^{2/3}}$$

Now P_{\oplus} is the period of a satellite at one Earth radius or

$$P_{\oplus} = \frac{2 \cdot \pi \cdot 3957}{5 \cdot 60 \cdot 60}$$

$$P_{\oplus} = 1.38 \text{ hr.} \quad \text{and}$$

$$P_{\Delta} = 24 \text{ hr} \quad \text{then}$$

$$a_{\Delta} = \left(\frac{24}{1.38} \right)^{2/3}$$

$a_{\Delta} = 6.8$ now the altitude, h , in miles is

$$h = (a_{\Delta} - 1) 3957$$

$$h = 23,000 \text{ miles}$$

Now the speed of the satellite is

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right)$$

but for a circular orbit $r = a$ therefore

$$\dot{s}^2 = \left(\frac{1}{r} \right)$$

$$\text{where } r = a_{\Delta} = 6.8$$

$$\mu = 1$$

$$\dot{s}^2 = 1 \left(\frac{1}{6.8} \right)$$

$$\dot{s}^2 = .147$$

$$\dot{s} = .384$$

Then the speed in miles/sec, u , is

$$u = \dot{s} \cdot 5$$

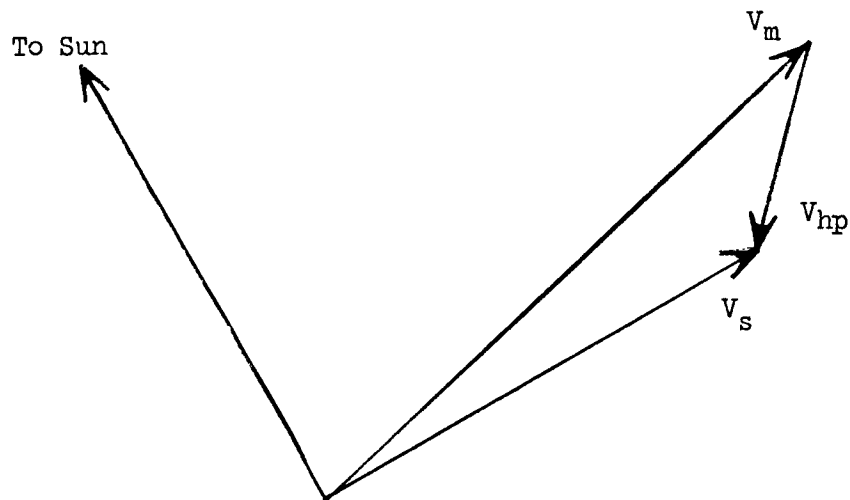
$$u = .384 \cdot 5$$

$$u = 1.92 \text{ miles/sec}$$

Planetary Approach

In designing missions to the planets the approach phase has a rather strong interaction with the entire mission, including the launch, transit, communication distance to Earth, and the flight time. To better understand this we will look specifically with the general problems associated with the planet Mars. These general concepts are applicable to all the planets with only minor modifications. Then to make the problems more tangeable we will consider the specifics of the Mars approach geometry as it will be during a 1971 opportunity.

The approach geometry at Mars is mainly determined by the magnitude and direction of the areocentric hyperbolic excess velocity. This velocity is the vectoral difference between the heliocentric velocity of Mars and the heliocentric velocity of the spacecraft at the Mars encounter, neglecting the gravitational influence of Mars on the spacecraft. This relationship is shown in Figure 1.



- V_m - heliocentric velocity of Mars
- V_s - heliocentric velocity of the Spacecraft
- V_{hp} - hyperbolic excess velocity.

Fig. 1

If we assume a simple coplanar Hohmann transfer between the Earth and Mars, we can obtain a quick estimate of the minimum value of V_{hp} . Using the geometry shown in Figure 2.

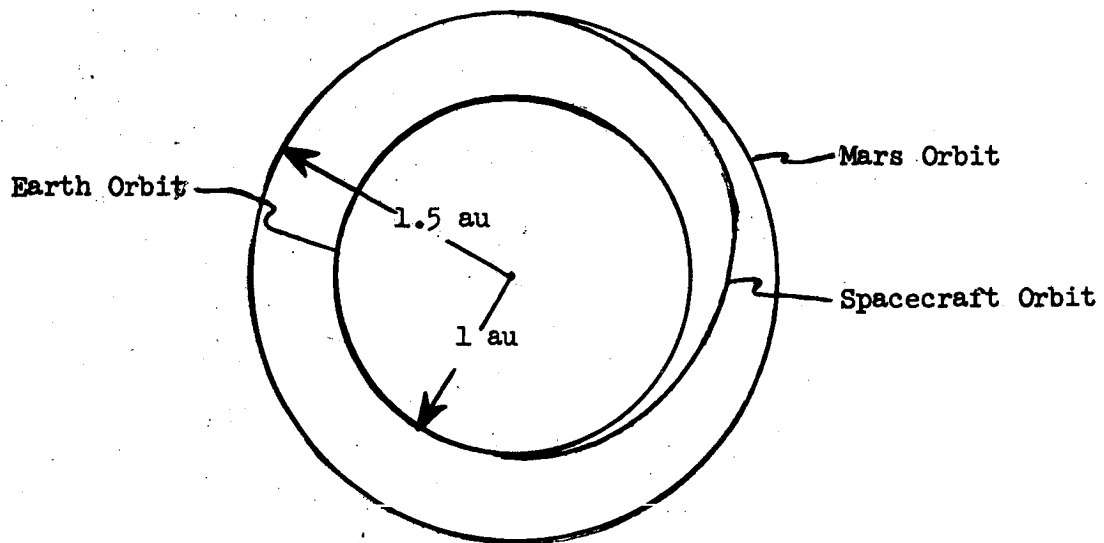


Fig. 2

First computing the velocity of the spacecraft v_s we have

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right)$$

$$a = \frac{1}{2} (1 + 1.5)$$

$$a = 1.25$$

$$r = 1.5$$

$$\mu = 1$$

$$\dot{s}^2 = 1 \left(\frac{2}{1.5} - \frac{1}{1.25} \right)$$

$$\dot{s} = .73$$

Now the heliocentric velocity of the Earth is approximately 30 km/sec, therefore

$$v_s = 30 \times .73$$

$$v_s = 21.90 \text{ km/sec}$$

and the heliocentric velocity of Mars is

$$\dot{s}^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right)$$

$$r = a = 1.5 \text{ then}$$

$$\dot{s}^2 = 1 \left(\frac{1}{1.5} \right)$$

$$\dot{s} = .815 \text{ and then}$$

$$v_m = 30 \times .815$$

$$v_m = 24.45 \text{ km/sec}$$

Now since we assumed a Hohmann transfer

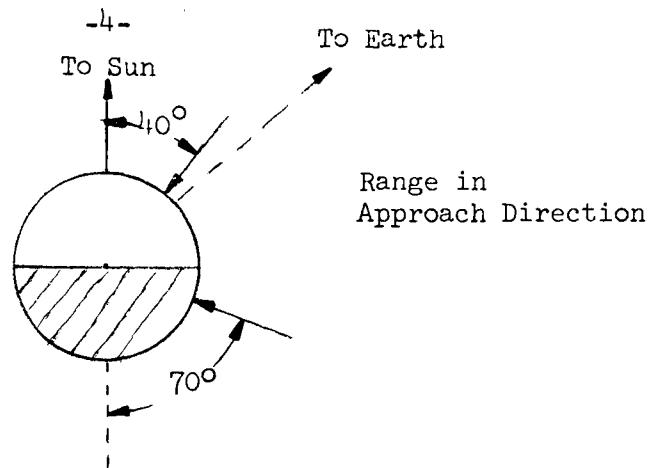
$$v_{hp} = v_m - v_s$$

$$v_{hp} = 24.45 - 21.90$$

$$v_{hp} = 2.55 \text{ km/sec}$$

It must be noted that this value is a minimum, assuming the orbits are circular; the transfer is co-planer and Hohmann. The actual minimum is 2.82 km/sec for 1971; 2.40 km/sec for 1973. By inspection of the vis-viva integral it can be seen that the spacecraft velocity at Mars will always be less than the planet's velocity, since the value of r is identical and the semi-major axis of the transfer will always be less than the semi-major axis of Mars, for transfers that are reasonably close to optimum.

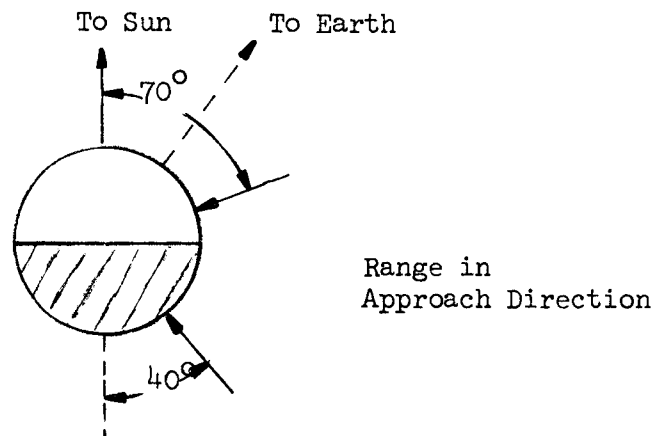
Two basic types of transfer trajectories must be considered, Type I and Type II. Type I trajectories have heliocentric transfer angles less than 180° , where the heliocentric transfer angle is measured from the position of the Earth at launch to the position of Mars at encounter. The Type I trajectories generally approach Mars from the lighted side, see fig. 3.



Type I Approach Direction

Fig. 3

Type II trajectories have heliocentric transfer angles greater than 180° , and approach Mars generally from the dark side, see fig. 4.



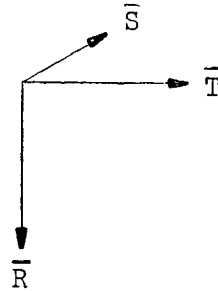
Type II Approach Direction

Fig. 4

The effects of approaching from the lighted or dark side of the planet can have some profound implication on such science experiments as television.

To better understand and visualize the entire geometry problem about Mars we will define a Mars centered coordinate system. This system will be a

right handed, three dimension, orthogonal system, see Fig. 5.



Approach Coordinate System

Fig. 5

This system is composed of three unit vectors, \bar{R} , \bar{S} , \bar{T} such that $\bar{R} = \bar{S} \times \bar{T}$. \bar{T} is parallel to the ecliptic, \bar{S} is parallel to the direction of the hyperbolic approach asymptote and \bar{R} completes the system. It is important to understand this system to make any progress in this entire approach problem.

Within this coordinate system we would like to know the location of the Earth and the Sun. This is important since we must communicate with the Earth and derive solar power from the Sun. We will first define an angle, ZAP, the angle between the Mars - Sun vector at encounter and the hyperbolic excess velocity vector. This angle is close to the Mars - spacecraft - Sun angle a few days before encounter. It should be noticed that if ZAP is less than 90° , the approach is from the dark side; for ZAP greater than 90° , the approach is from the lighted side. Another important angle ETS is also defined. ETS is the angle measured clockwise from the \bar{T} axis to the negative projection of the Mars - Sun vector onto the $\bar{R} - \bar{T}$ plane. These two angles are shown in Figure 6.

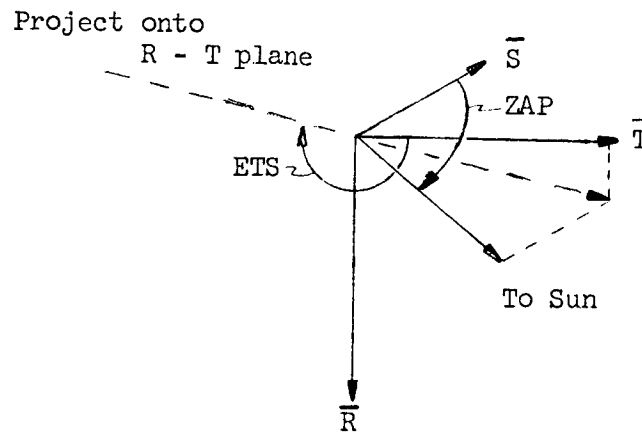


Fig. 6

A similar pair of angles ZAE and ETE are defined for the Earth. ZAE is the angle between the Mars - Earth vector and the hyperbolic excess velocity vector. ETE is the angle measured clockwise from the T axis to the negative projection of the Mars Earth vector onto the $R - T$ plane. The reason for measuring these angles (ETE and ETS) to the negative projections will become apparent.

To determine where the spacecraft actually flies by the planet in this coordinate system we will define a point in the $R - T$ plane where the hyperbolic approach asymptote passes through that plane. This point will be defined as the aiming point, and the $R - T$ plane the aiming plane. The aiming point can then be defined by the vector \bar{B} which has magnitude $|\bar{B}|$ and orientation θ to the T axis or by the components of \bar{B} , $(\bar{B} \cdot \bar{T})$ and $(\bar{B} \cdot \bar{R})$ as shown in figure 7 and 8.

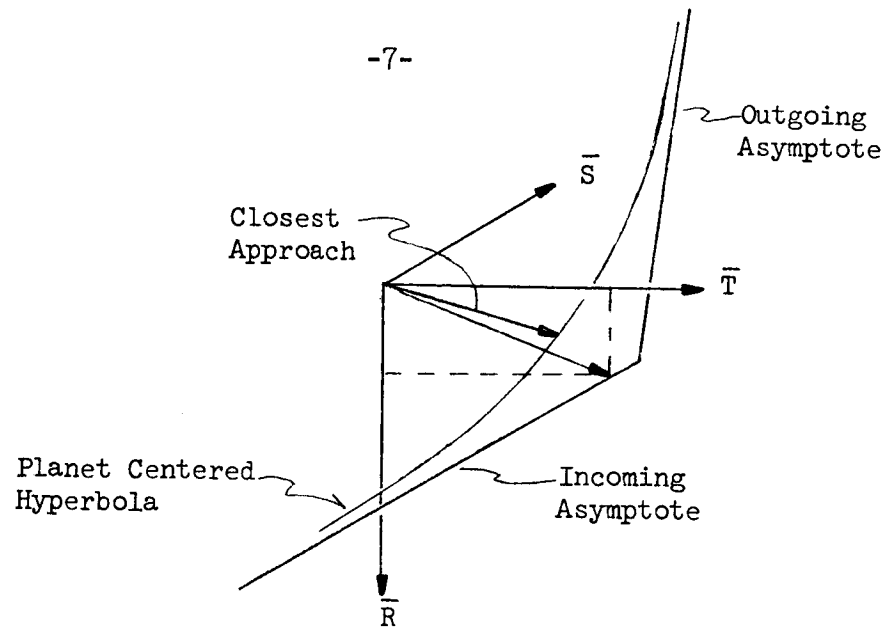
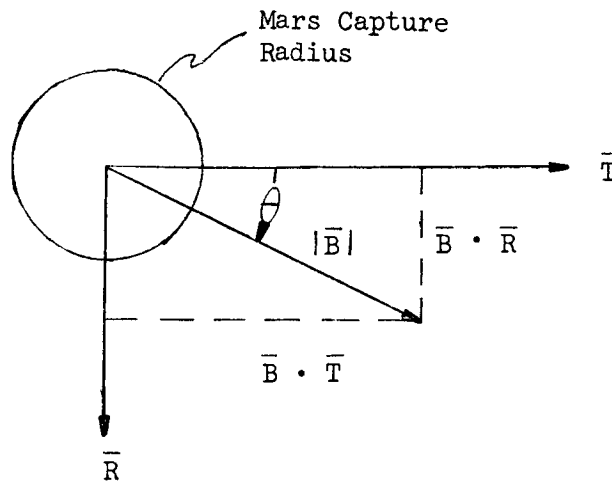


Fig. 7

For clarity the aiming plane is shown in fig. 8.



Aiming Plane

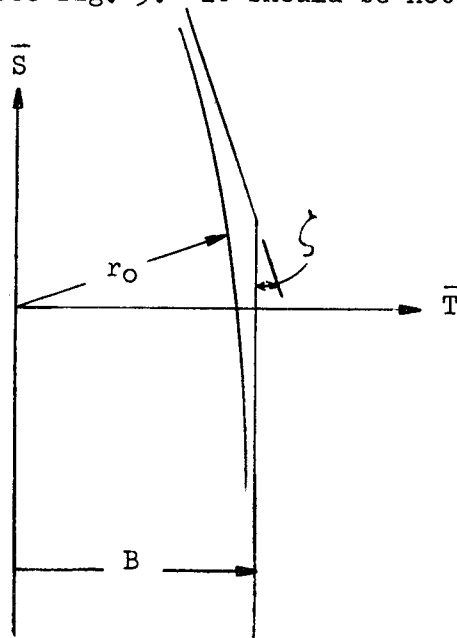
Fig. 8

Now with this definition the reason for measuring ETE and ETS to the negative projections becomes apparent. If $\theta = \text{ETE}$ then the Earth as seen by the spacecraft will be occulted and if $\theta = \text{ETS}$ the Sun as seen by the spacecraft

will be occulted. The time after encounter at which these occultations will occur depend on the magnitude of B and the hyperbolic excess velocity. In the case of an orbiter Earth occultation will occur at ETE and $ETE + \pi$, similarly for ETS.

Now with this concept of targeting, or aiming, the approach asymptote in the $R - T$ plane at a massless planet it becomes easy to transfer from a heliocentric orbit to an areocentric orbit. This concept also allows the approach phase to be treated independently from the interplanetary phase. We can now investigate what the real near planet geometry is given the miss parameter and the hyperbolic excess velocity, v_{∞} .

The first parameter of interest is the radius of closest approach, r_0 . This is the shortest perpendicular distance from the actual flyby trajectory to the center of the planet, see fig. 9. It should be noted in fig. 9 that \bar{T}



Miss Parameter B , Closest Approach r_0

Fig. 9

is in the plane of the paper for $\theta = 0$ only. The relationship between r_0 and B is

$$B = r_o \sqrt{1 + \frac{2 GM}{r_o v_{oo}^2}}$$

where for Mars $GM = 4.298 \times 10^4 \text{ km}^3/\text{sec}^2$

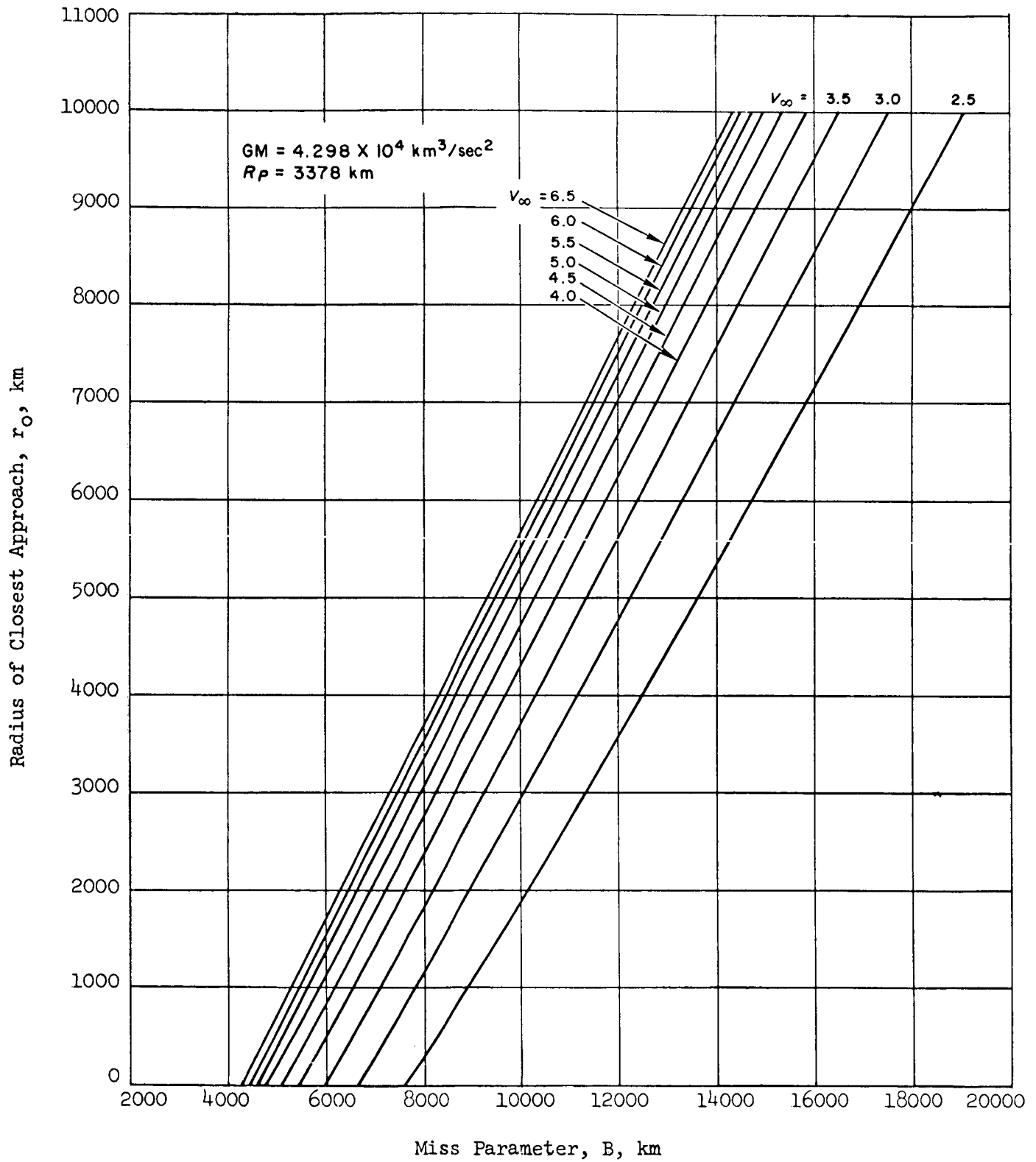
this is shown in fig. 10 for different values of v_{oo} .

The bending angle ζ as shown in fig. 9 is plotted in fig. 11, again verses miss parameter, for different values of v_{oo} .

The entry angle, θ_e , and range angle, ϵ , are two important parameters in considering the atmospheric entry problems. These parameters are shown in fig. 12 and plotted in fig. 13.

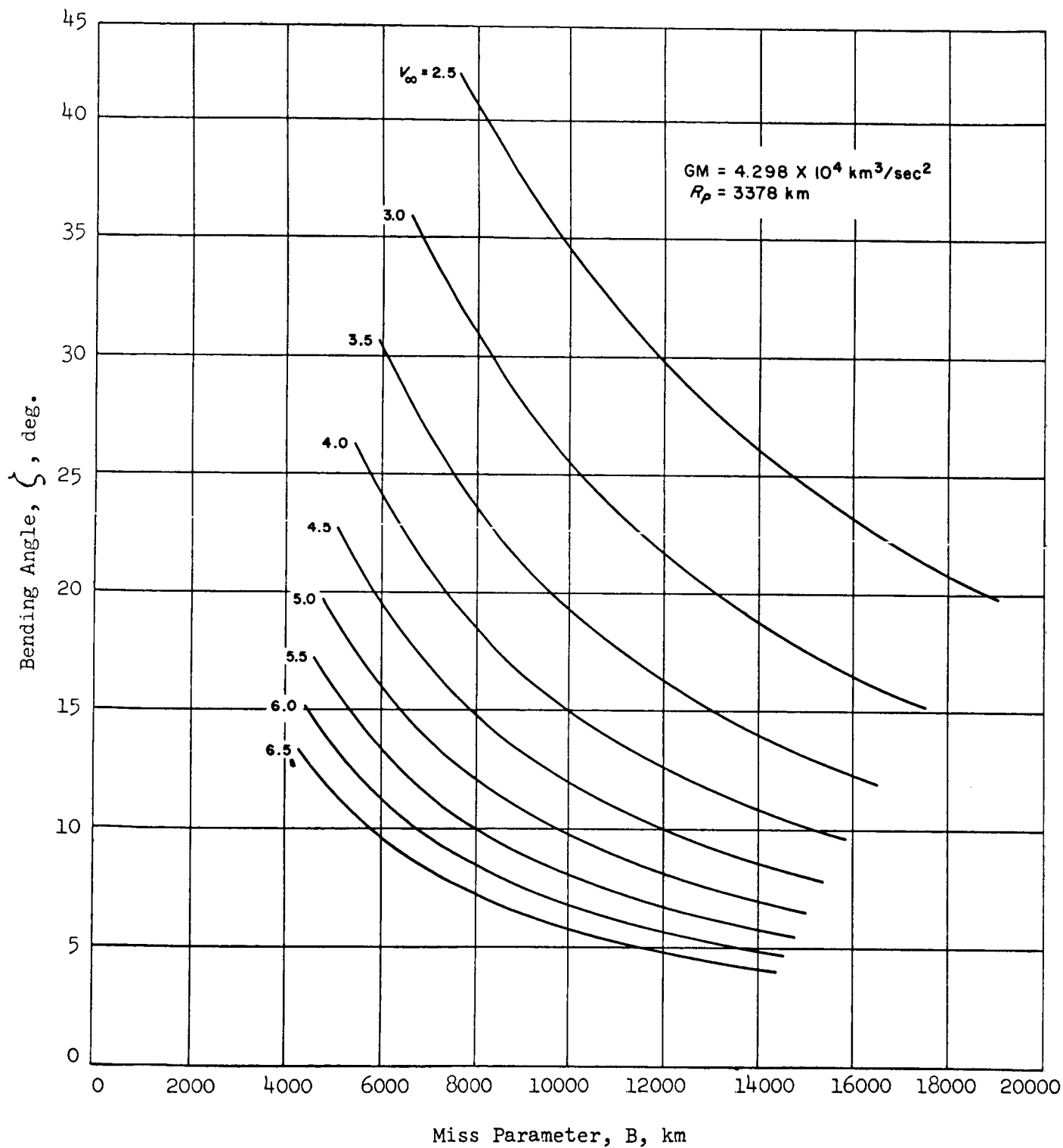
In addition to changing the shape of the approach trajectory the gravitational effect of the planet on the spacecraft also increases the hyperbolic excess velocity, v_h . The hyperbolic excess velocity is v_{oo} at very large distances from the planet ($R = \infty$) and v_e at the upper atmosphere. This entry velocity is shown in fig. 14. The hyperbolic excess velocity is related to v_{oo} as shown below

$$v_h^2 = v_{oo}^2 + \frac{2 GM}{R}$$



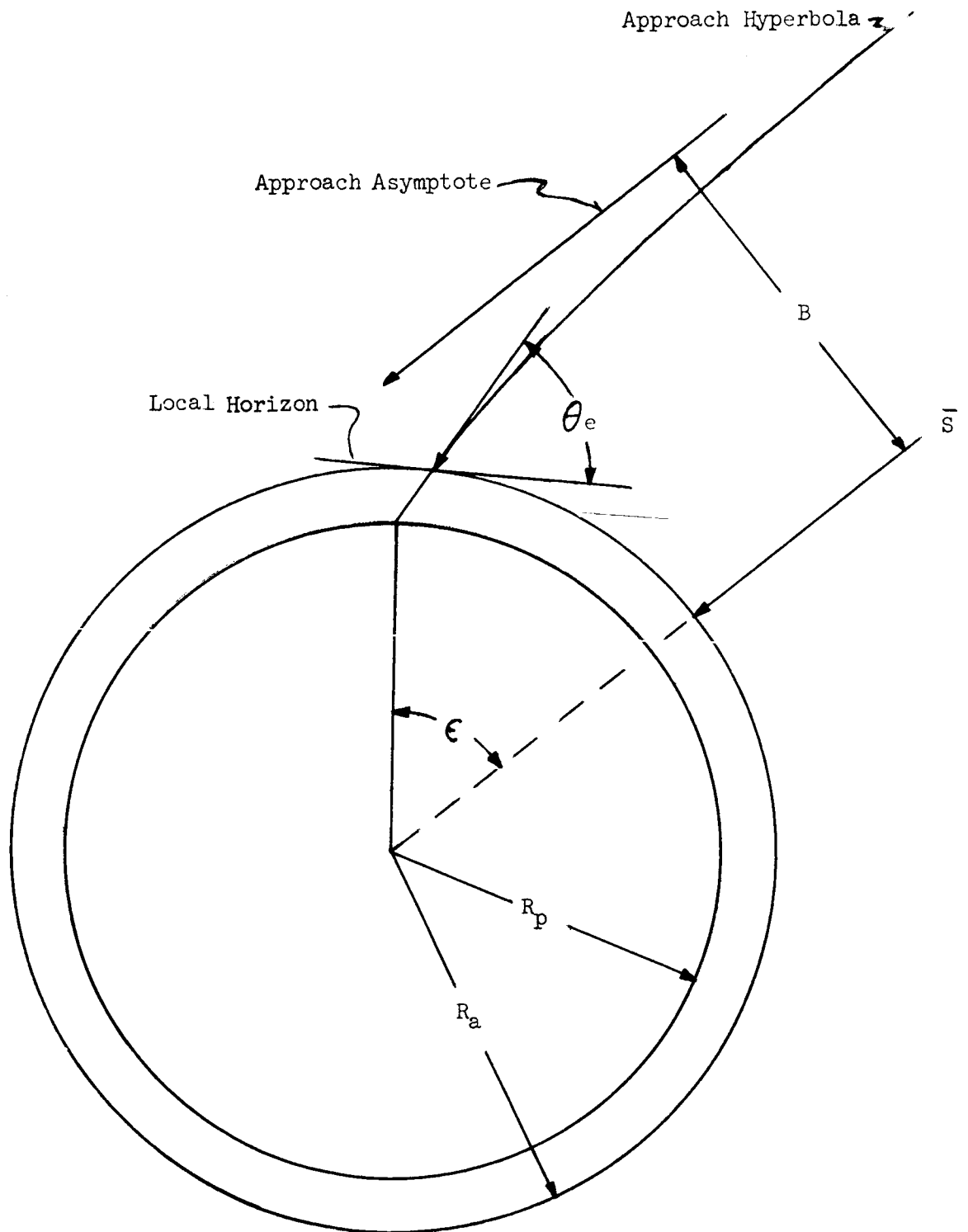
Radius of Closest Approach vs. Miss Parameter

Fig. 10



Bending Angle vs. Miss Parameter

Fig. 11

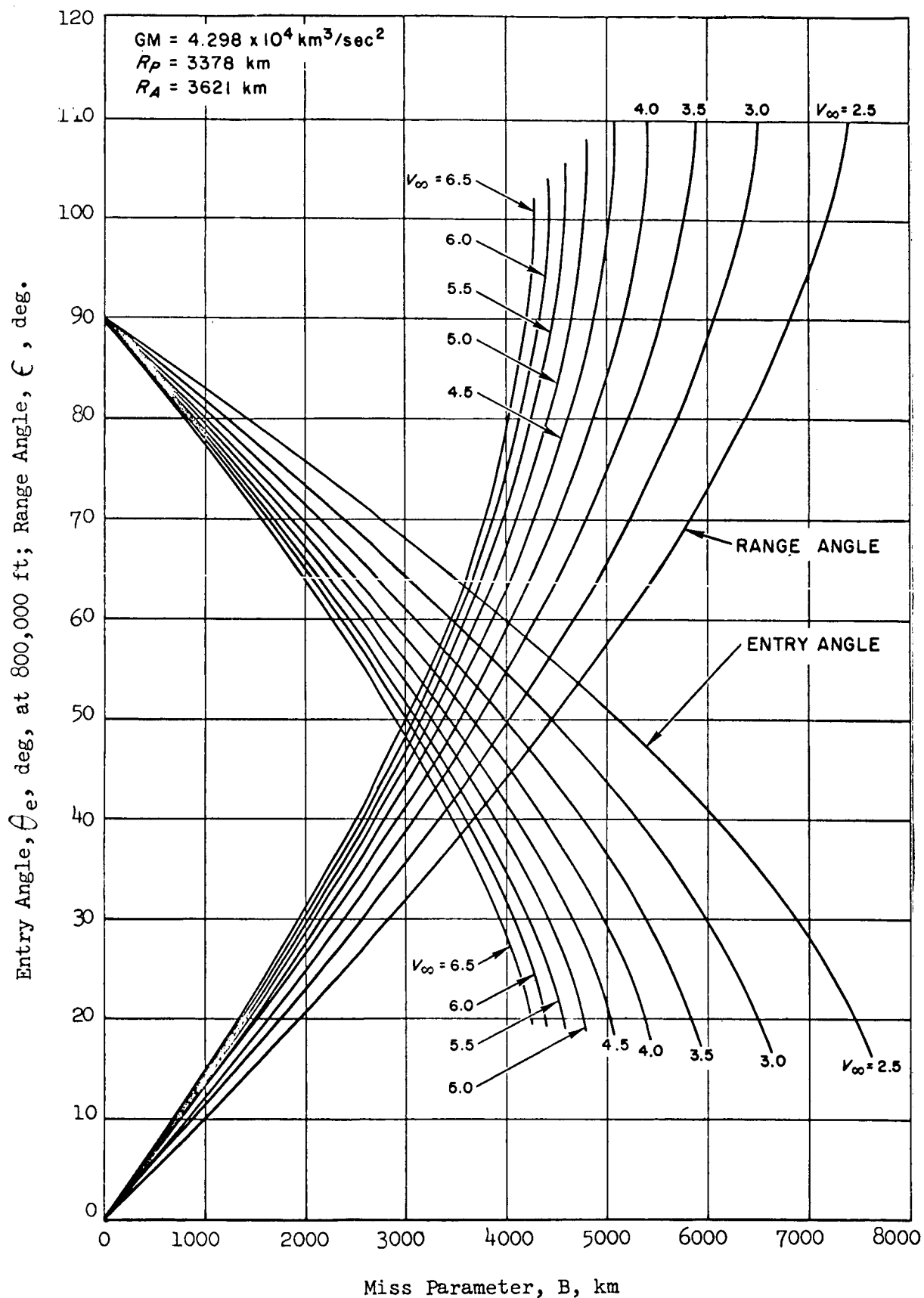


$$R_p = 3378 \text{ km}$$

$$R_a = 3621 \text{ km}$$

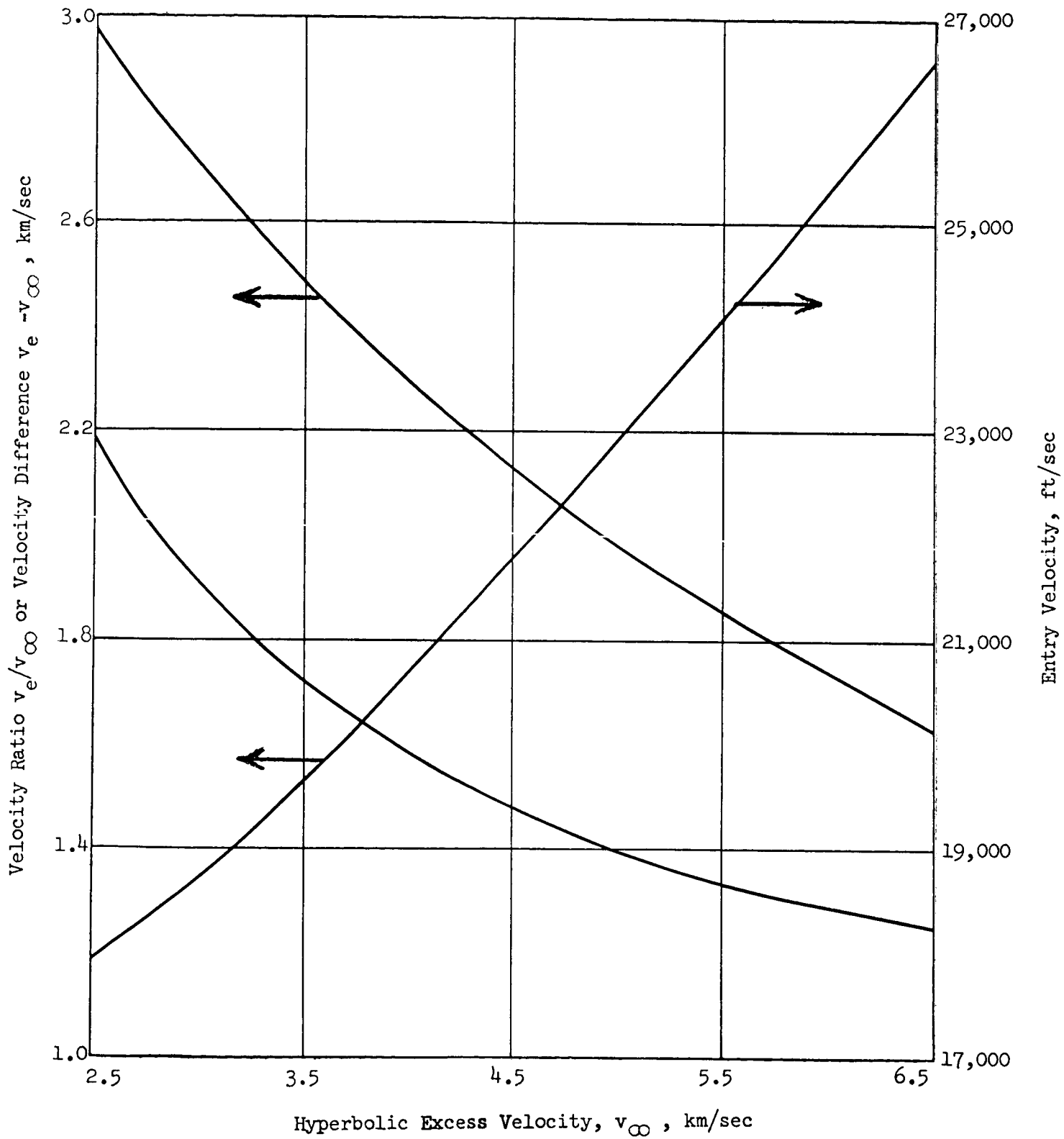
Range Angle, ϵ , and Entry Angle θ

Fig. 12



Entry Angle & Range Angle vs. Miss Parameter

Fig. 13

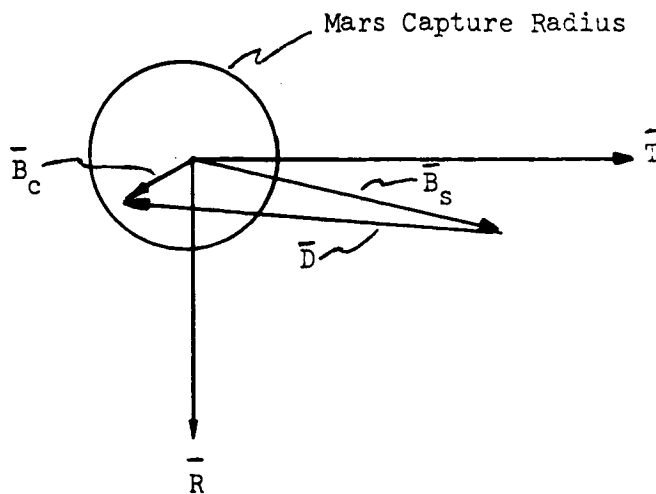


Entry Velocity vs. Hyperbolic Excess Velocity

Fig. 14

It is possible now to consider the problems associated with separating a capsule from a spacecraft in the vicinity of the planet. One could consider a mission in which a small entry capsule was carried by a spacecraft to Mars then separated from the spacecraft and placed on an impact trajectory to the planet. The spacecraft could then serve as a relay station between the capsule and the Earth, or the capsule could transmit the information directly back to Earth. In either or both cases it may be desirable for the spacecraft to perform some experiments, such as television, when it flies by the planet. Thus we see that there exists some rather important geometrical relations between the planet, the spacecraft, the capsule and the Earth.

First let us look at the magnitude and direction of the maneuver required to place the capsule on an impact trajectory. If the spacecraft is targeted at some aiming point in the R - T plane with a value \bar{B}_s and it is desirable to have the capsule targeted to an aiming point in the R - T plane with a value \bar{B}_c then there exists a deflection distance, \bar{D} as shown in fig. 15



Spacecraft - Capsule Aim Diagram

Fig. 15

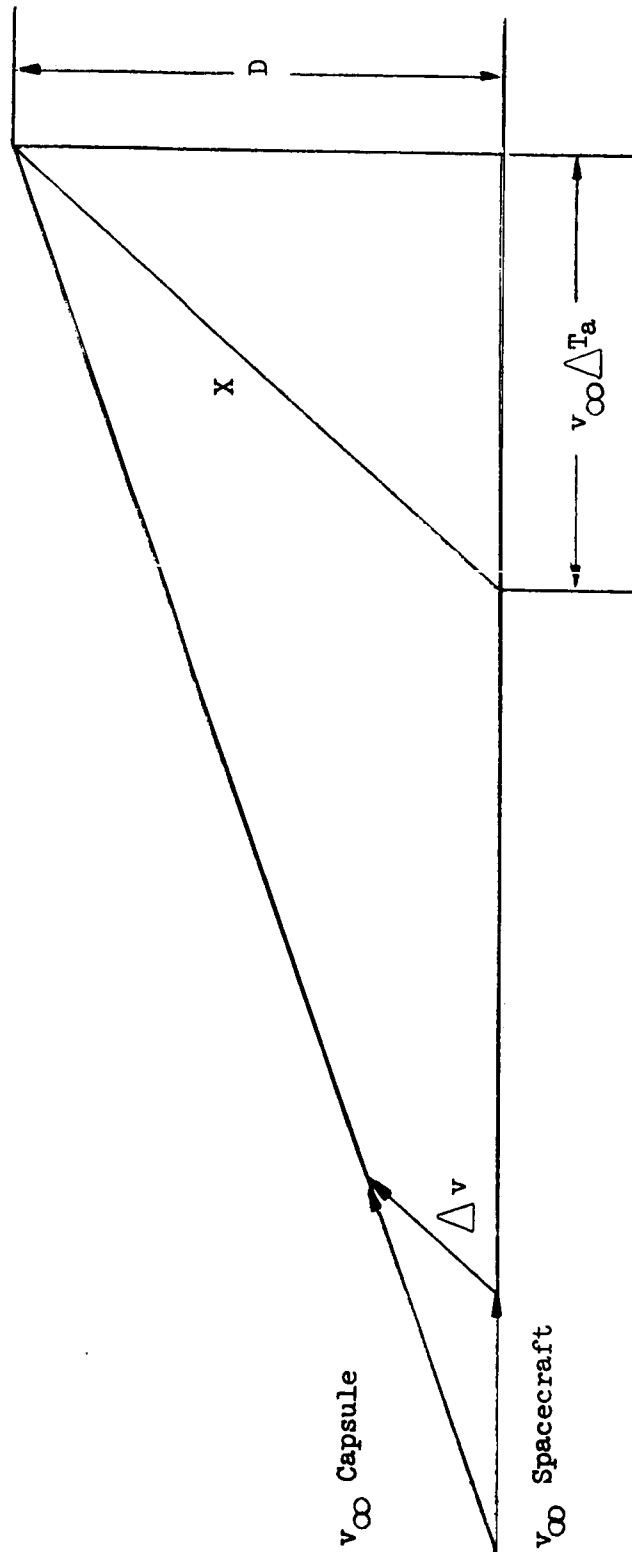
From figure 15 it can be seen that $\bar{D} = \bar{B}_c - \bar{B}_s$. The spacecraft approach asymptote \bar{S}_s and \bar{D} form a plane. This plane is perpendicular to the R - T plane and shall be called the maneuver plane. Now using the concept of this maneuver plane it is possible to determine the magnitude and direction of the required capsule deflection maneuver. Another important parameter is the difference in arrival time at the planet between the capsule and the spacecraft. For example if the capsule is transmitting to the spacecraft then it is important for the capsule to arrive at the planet before the spacecraft. This difference in arrival time, ΔT_a , places some important considerations on the mission design as will be seen later.

The time before the spacecraft encounter at which the maneuver is made is T_f . With these parameters a good approximation to the actual geometry can be made assuming the motions to be rectilinear. This is shown in Fig. 16, where the plane of the paper corresponds to the maneuver plane. The communication range at encounter, X, is given by

$$X = \sqrt{D^2 + (v_{\infty} \Delta T_a)^2}$$

where $v_{\infty} \Delta T_a$ is the distance the spacecraft moves after the capsule impacts. These three parameters, X, D and $v_{\infty} \Delta T_a$, uniquely determine the desired encounter geometry. The velocity increment, Δv , applied to the capsule has two components

$$\begin{aligned} v_1 &= \Delta v \cos & \text{and} \\ v_2 &= \Delta v \sin & \text{such that} \\ v_1 T_f &= v_{\infty} \Delta T_a & \text{and} \\ v_2 T_f &= D & \text{and since} \\ \tan \alpha &= \frac{v_2}{v_1} & \text{then} \end{aligned}$$



Maneuver Plane Geometry

Fig. 16

$$\tan \alpha = \frac{D}{v_{\infty} \Delta T_a} \quad \text{therefore}$$

$$\Delta v T_f = X \quad \text{or}$$

$$\Delta v = \frac{X}{T_f} .$$

Thus the required velocity increment is seen to vary inversely with the separation time and directly with the communication distance, X , at encounter, and the application angle, α , is given by

$$\alpha = \tan^{-1} \left(\frac{D}{v_{\infty} \Delta T_a} \right)$$

In referring to figure 16 care should be taken to realize that the actual geometry is significantly different since the v_{∞} of the capsule and spacecraft are almost parallel and have essentially the same magnitude.

Some first order approximations to the accuracy of such a maneuver can be made as follows. There are two main error sources to be considered, σ_v , the error in the total velocity increment, and σ_p the error in pointing (or direction), about two orthogonal axis. The velocity error will be assumed to be a fixed percentage of the total magnitude and the error in pointing an absolute value. These two errors map into an aiming point error with three orthogonal components, two in the aiming plane and one normal to. The two "in plane" errors have directions along \bar{S} and \bar{D} ; the "out of plane" error is normal to \bar{S} and \bar{D} . These three errors will be defined as follows:

σ_{in} - the in plane error in the D direction,

σ_s - the in plane error in the S direction,

σ_{out} - the out of plane error

The value of σ_{in} is given by

$$\sigma_{in} = \sqrt{(\sigma_v D)^2 + (\sigma_p v_{\infty} \Delta T_a)^2}$$

This can be seen in figure 17 and 18 and the following derivation.

Both the pointing and velocity errors map into position errors as shown in fig. 17 and 18. From similar triangles it can be seen that

$$\frac{\sigma_v X}{X} = \frac{\sigma_1}{D} \quad \text{or}$$

$$\sigma_1 = \sigma_v D \quad \text{and}$$

$$\frac{\sigma_p X}{X} = \frac{\sigma_2}{v_{\infty} \Delta T_a} \quad \text{or}$$

$$\sigma_2 = \sigma_p v_{\infty} \Delta T_a$$

Now the total error in the D direction is the RSS of these two or

$$\sigma_{in} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

$$\sigma_{in} = \sqrt{(\sigma_v D)^2 + (\sigma_p v_{\infty} \Delta T_a)^2}$$

Similarly the inplane error in the \bar{S} direction is

$$\sigma_s = \sqrt{(\sigma_p D)^2 + (\sigma_v v_{\infty} \Delta T_a)^2}$$

$$\frac{\sigma_p X}{X} = \frac{\sigma_3}{D} \quad \text{or}$$

$$\sigma_3 = \sigma_p D \quad \text{and}$$

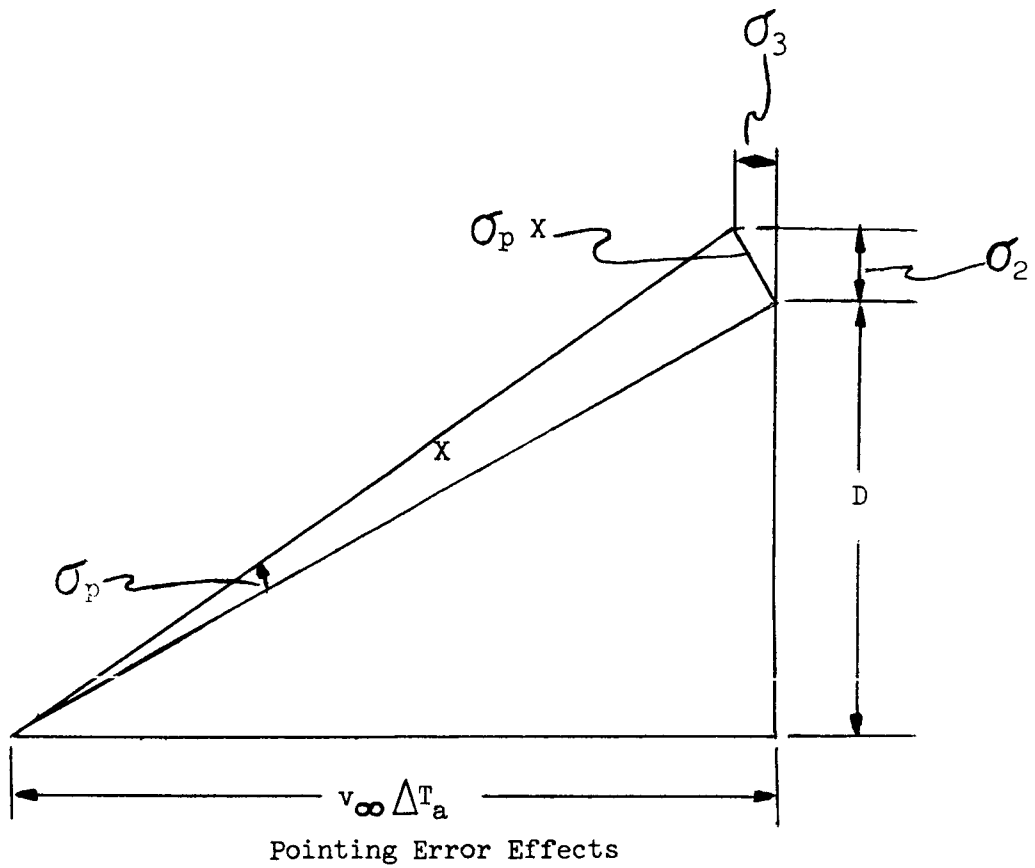


Fig. 17

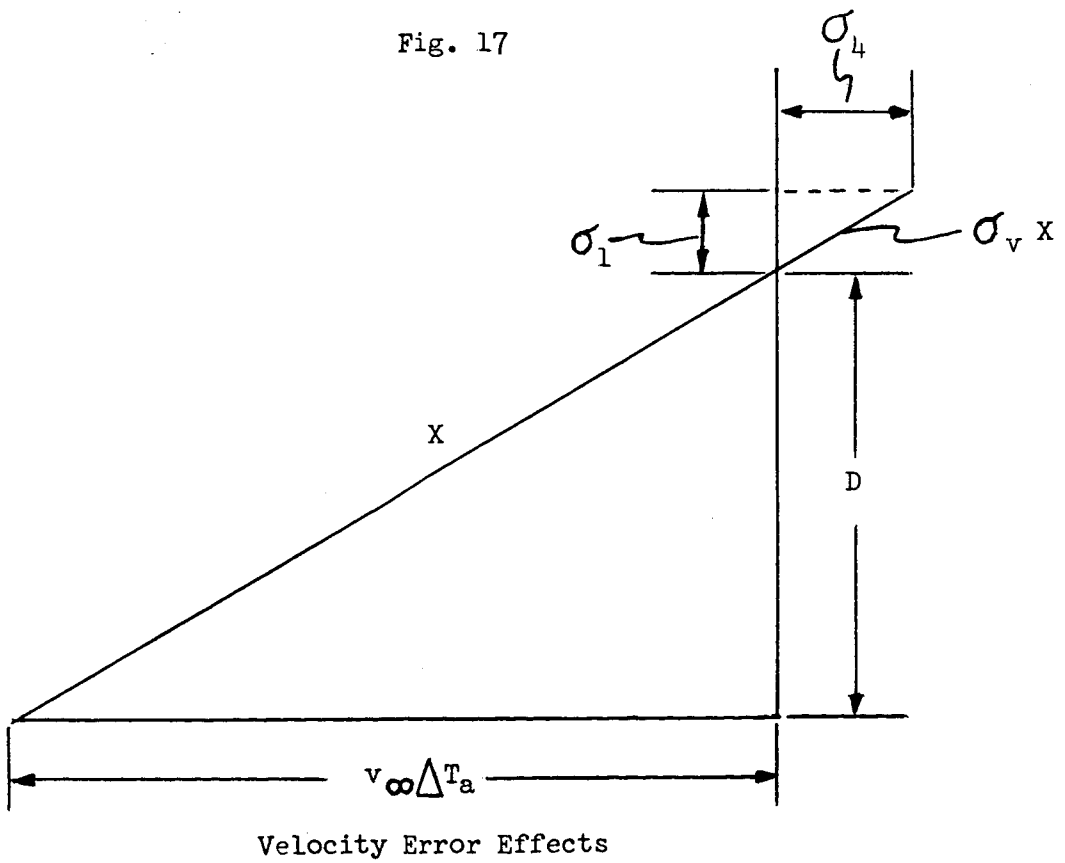


Fig. 18

$$\frac{\sigma_v X}{X} = \frac{\sigma_4}{v_\infty \Delta T_a} \quad \text{or}$$

$$\sigma_4 = \sigma_v v_\infty \Delta T_a \quad \text{and then}$$

$$\sigma_s = \sqrt{\sigma_3^2 + \sigma_4^2} \quad \text{or}$$

$$\sigma_s = \sqrt{(\sigma_p D)^2 + (\sigma_v v_\infty \Delta T_a)^2}$$

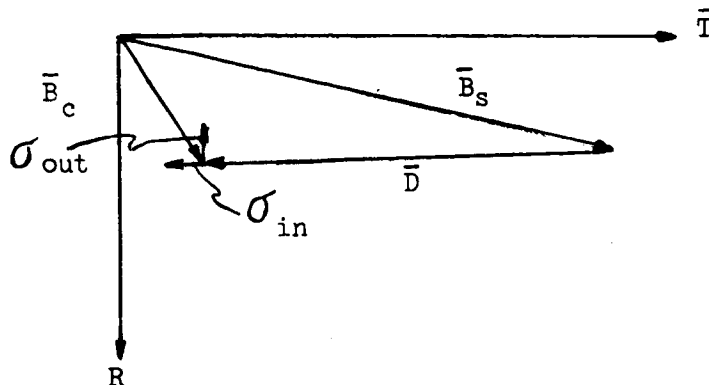
And finally the component out of the plane is simply

$$\sigma_{out} = \sigma_p X$$

The in plane error in the S direction is important in determining the error in arrival time, σ_t , where

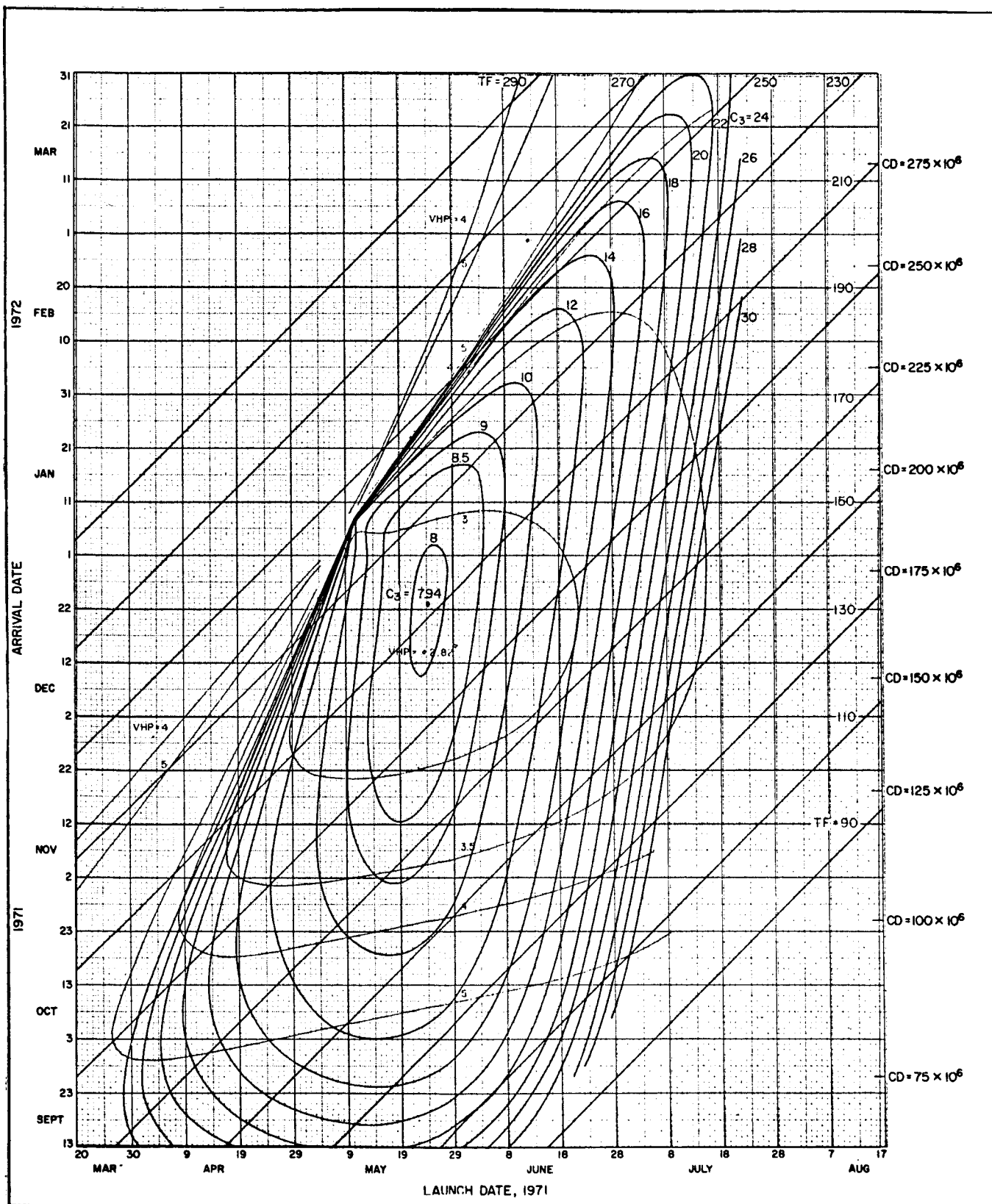
$$\sigma_t = \frac{\sigma_s}{v_\infty}$$

It should be noted that these errors are only the relative errors between the spacecraft and the capsule. In addition to these there are orbit determination errors which account for the uncertainty of where the fly by trajectory is with respect to the actual position of the planet. This error must be add (RSS) to these errors. The σ_{in} and σ_{out} then map in the R - T plane as shown in fig. 19.



In Plane & Out of Plane Excitation Errors

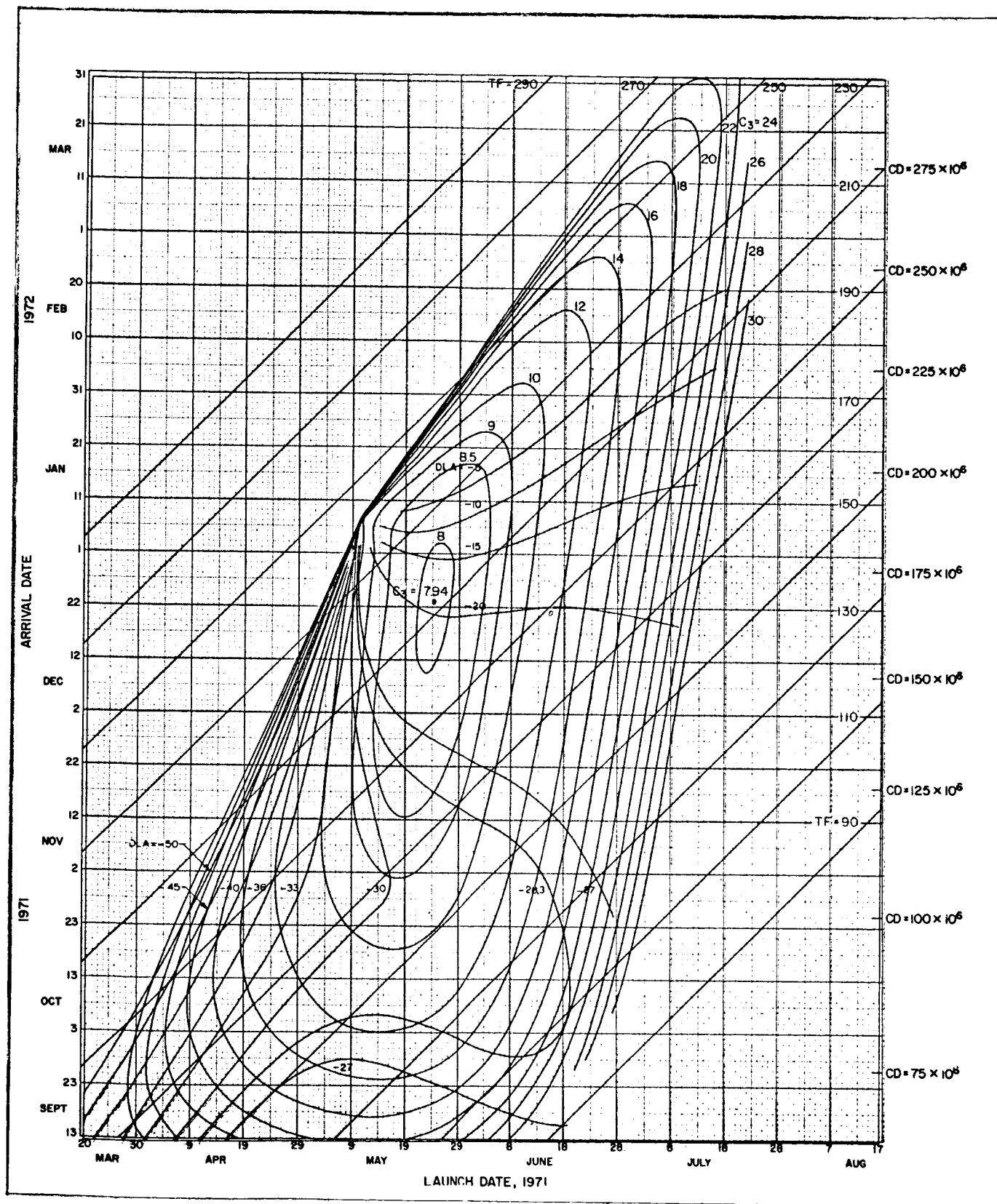
Figures 20 through 26 are basic design charts for the 1971 Type 1 trajectories. The closed containers are values of C_3 , which can be directly related to the payload capability of any launch vehicle.



Basic Trajectory Design Chart, 1971, Type 1

Hyperbolic Excess Velocity Relative to Mars

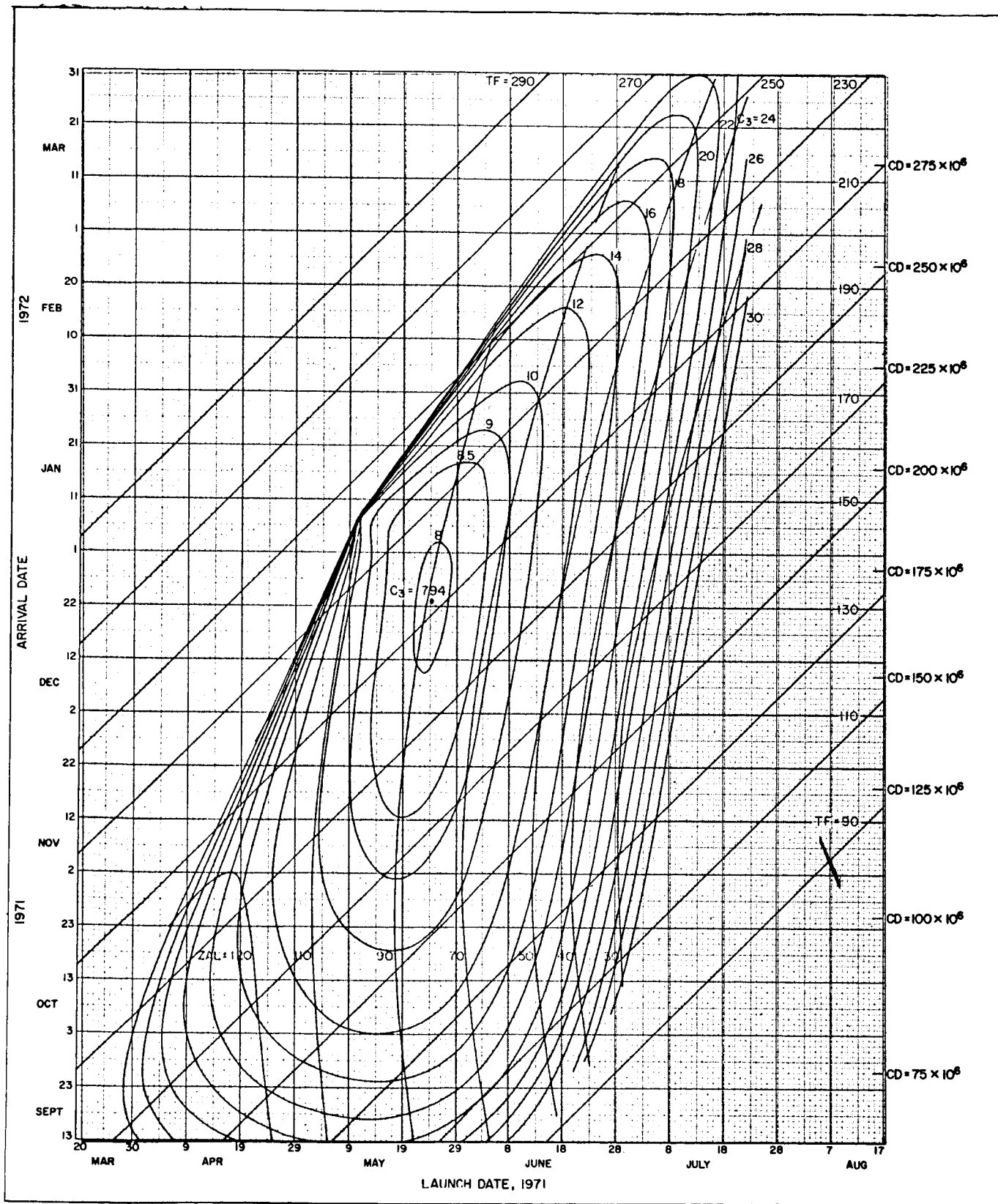
Fig. 20



Basic Trajectory Design Chart, 1971, Type 1

Declination of the Geocentric Departure Asymptote

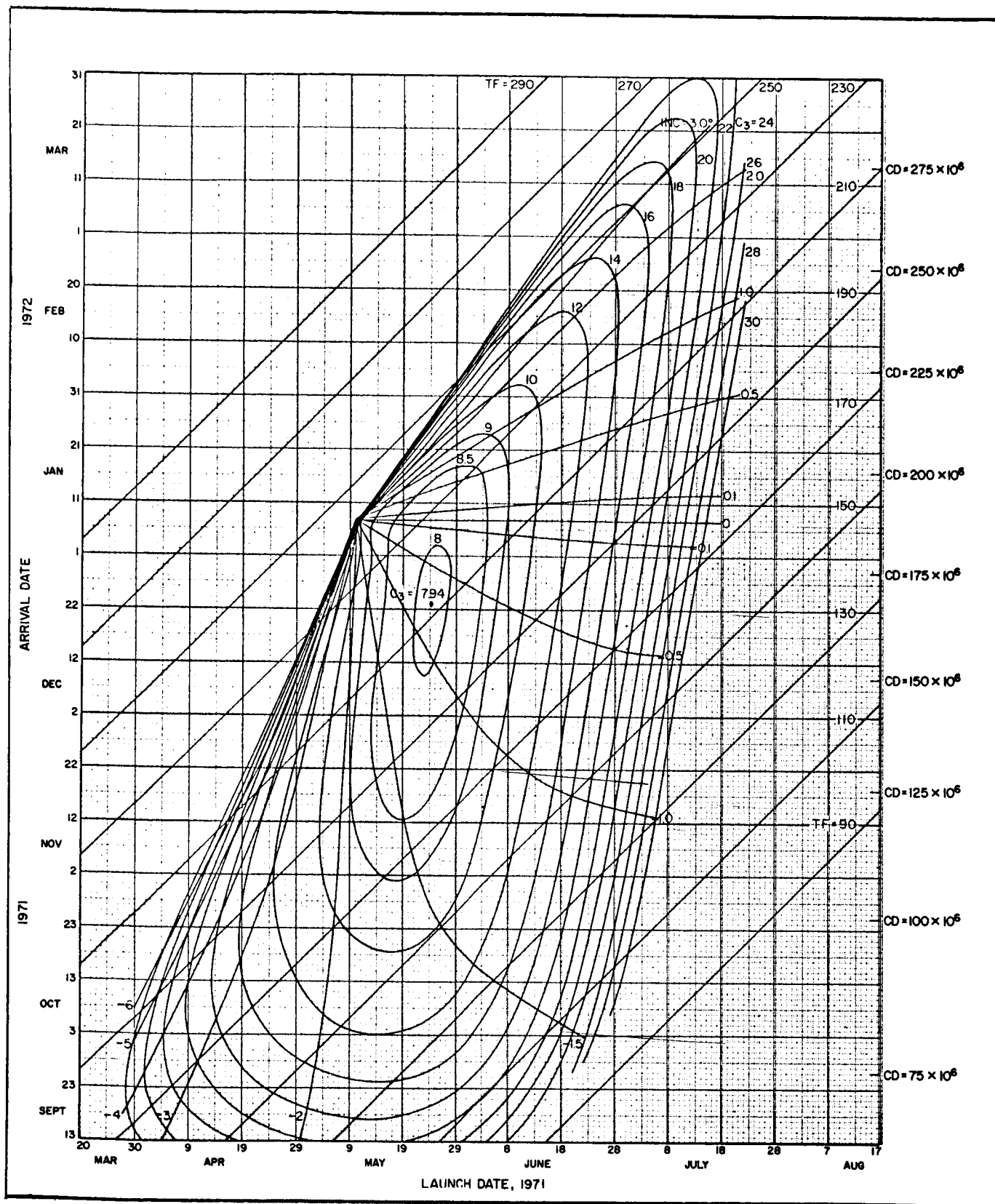
Fig. 21



Basic Trajectory Design Chart, 1971, Type 1

Angle Between the Sun-Earth Vector and Departure Asymptote

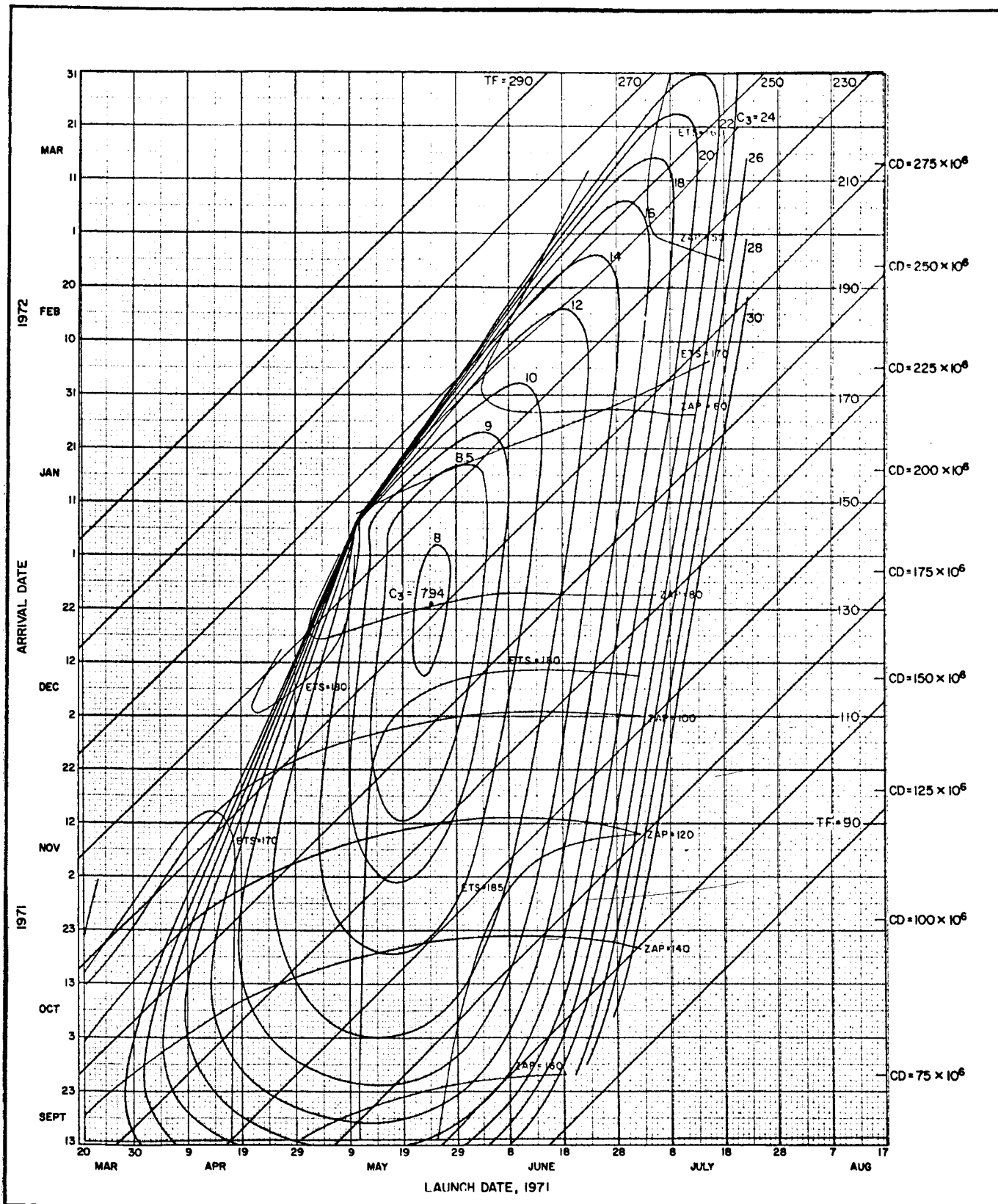
Fig. 22



Basic Trajectory Design Chart, 1971, Type 1

Inclination of the Transfer Plane to the Ecliptic Plane

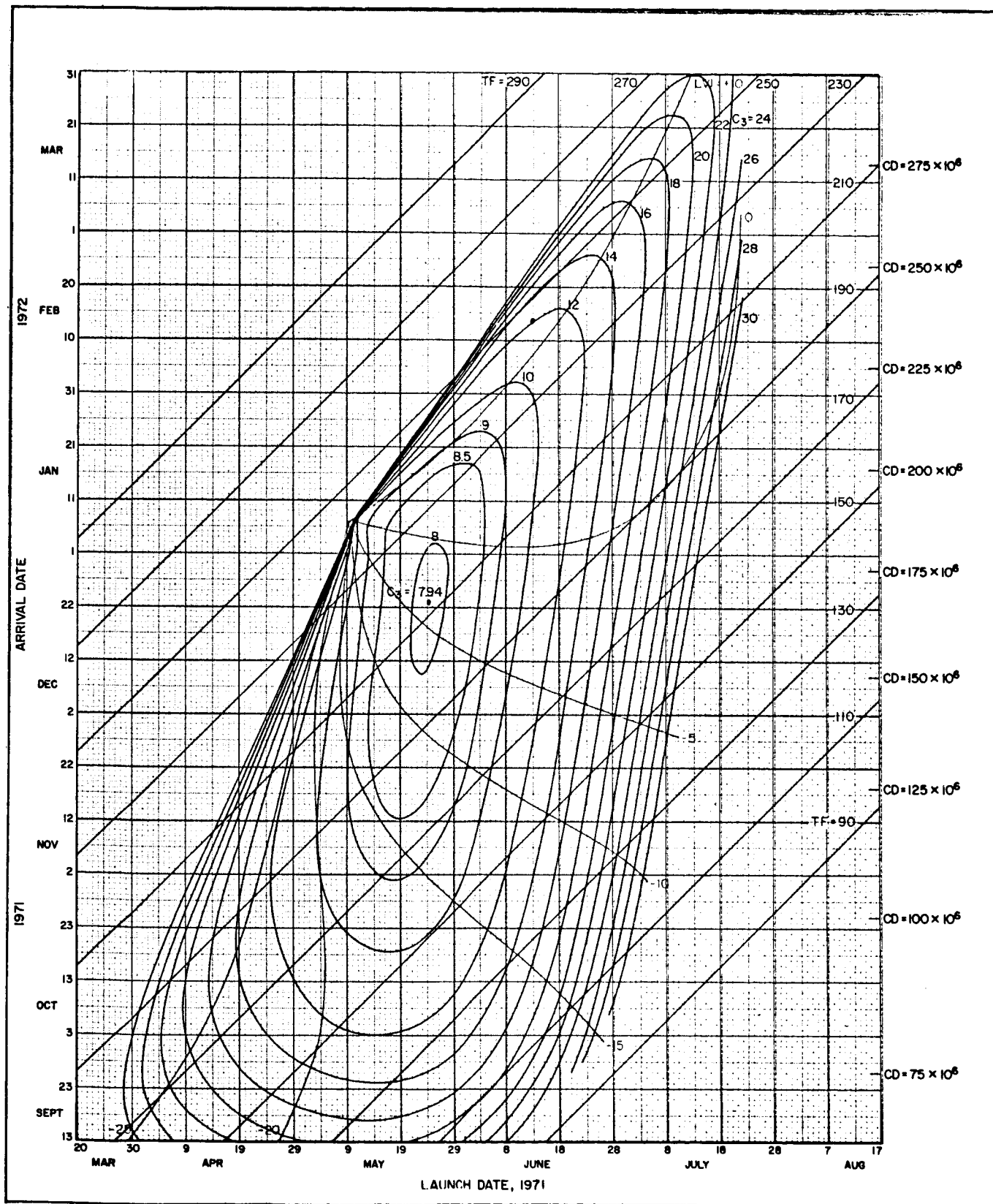
Fig. 23



Basic Trajectory Design Chart, 1971, Type 1

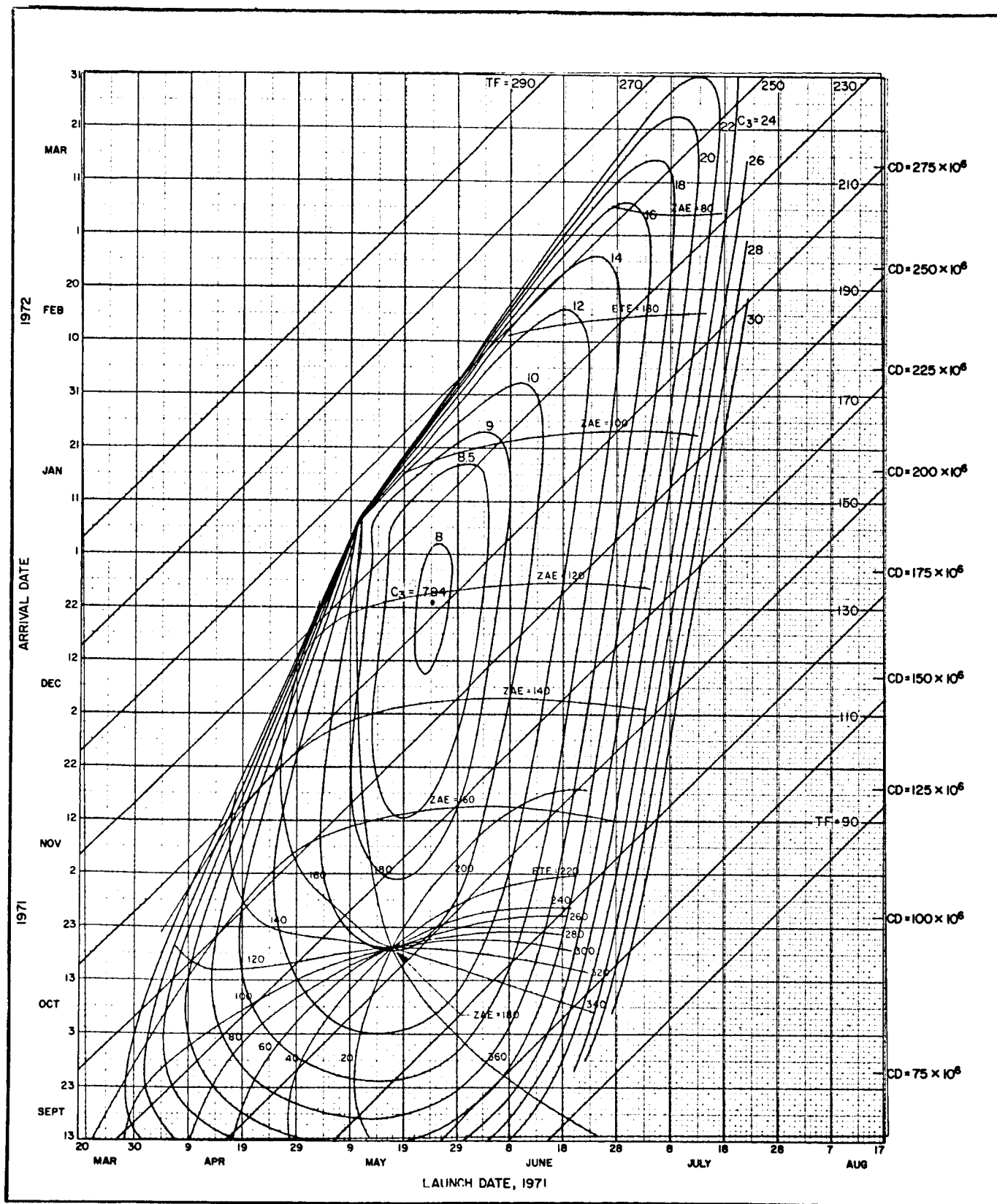
ZAP and ETS

Fig. 24



Basic Trajectory Design Chart, 1971, Type 1

Latitude of the Vertical Impact Point on Mars



Basic Trajectory Design Chart, 1971, Type 1

ZAE and ETE

Fig. 26

128
N 6 7. 8. 0. 4 6. 4

SPACE SCIENCE -

III

by

Roman K. C. Johns

PREFACE

The basic text for the Space Science course is, Sourcebook on the Space Sciences, edited by S. Glasstone, D. Van Nostrand Company, Inc., Princeton, New Jersey, 1965. The intention of the author of these notes is to supplement the basic text rather than to deal with matters expounded in the book. Some theoretical aspects are amplified and recent advancements in geophysics are discussed.

June, 1966

Roman K. C. Johns

Physics Department
Loyola University
Los Angeles, California

TABLE OF CONTENTS

Chapter	Title	Page
I	The Reference Ellipsoid	I-1
II	The Earth's External Gravitational Field	II-1
III	The Earth's Magnetic Field	III-1
IV	Radiative Processes in the Atmosphere	IV-1
	Bibliography and Periodicals	

I

THE REFERENCE ELLIPSOID

The shape of the Earth is of interest from the scientific point of view as well as from the practical point of view. Recent advancements in technology and instrumentation have in turn created increased demands concerning the extent and accuracy of the knowledge of the Earth's figure and its gravitational field.

A precise definition of the figure of the solid Earth is a difficult concept requiring differentiation between topography and crust; a more convenient concept is to define the figure of the Earth as consisting of its sea level. Of course, by this surface is understood the ocean surface formed only by the gravipotential of the rotating Earth, and not perturbed by winds, tides, local topography, and the like.

It must be pointed out that fundamentally there is no need for a reference surface; however, the complexity of the Earth's physical surface raises the question of selecting a geometrical figure which can serve as an adequate approximation and can be suitable for geometrical and mathematical operations. A surface of reference is primarily a matter of convenience for the three dimensional representation of relative locations below, on, or above the Earth's surface, and for performing mathematical computations. An adequate surface of reference has been found to be an ellipsoid.

1. The Figure of the Earth

A brief historical review of problems related to the determination of the Earth are given in the following article, "The Figure of the Earth," by R. K. C. Johns which appeared in the Journal of the Royal Astronomical Society of Canada, Vol. 53, pp. 257-263, (1959).

The conception of what the earth looks like and its position in the universe has varied through the ages. The interest in the figure of the earth also has been motivated by practical considerations, or in order to travel and navigate from one place to another, the directions and the distances must be known.

The first approximate but scientific notion of the shape or figure of the earth was a sphere and was given by Aristotle and Eratosthenes. Twenty centuries later Newton calculated that centrifugal force causes the earth to bulge at the equator and thus showed the earth to be an oblate spheroid. Now we are learning more about the actual shape of the earth and are hopeful of obtaining more accurate information regarding the relative positions of continents. It has become apparent that man-made satellites provide an important tool in the hands of geodesists.

In ancient civilizations the earth was considered to be a flat disk surrounded by oceans and in one way or another placed in the centre of the celestial system. For example, early Chinese maps show the lands of the earth consisting of islands, swimming in water, surrounding one enormous country—the middle Kingdom of China.

The Greek philosopher Plato was the first to have had the courage to assert that the earth is not the centre of the universe but only one of many planets. After him Aristotle suggested the earth is a sphere. The Greeks were able to measure the obliquity of the earth's orbit. In the third century B. C., Eratosthenes, the librarian of Alexandria, completed the first determination of the earth's radius in human history. He chose two stations, one in Aswan and the other in Alexandria. At the summer solstice in Aswan the sun at noon time was exactly overhead shining straight in a deep well. At the same time in Alexandria, 500 miles north, the sun's rays and the zenith were enclosing an angle, z , a little over 7 degree of arc. (See figure 1.1)

Assuming that the sun's rays are parallel, the angle measured at Alexandria corresponds to the angle between the plumb lines which is identical with the angle at the centre of the earth's sphere, as indicated in figure 1. Eratosthenes had exceptionally good luck. His results were nearly perfect, although everything he did seems to have been inaccurate. His arc was measured incorrectly; so was his angle. Eratosthenes measured the distance between the two stations, calculating a radius for the earth of 3,488 nautical miles. A recent figure calculated by the U.S. Army Map Service is 3,444 nautical miles.

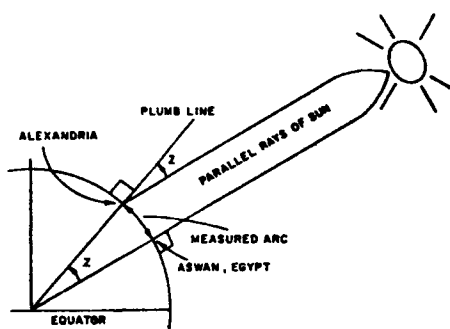


Fig. 1.1-How Eratosthenes measured the earth.

Through Arab scientists Greek geodetic knowledge reached Renaissance Italy from which scientific interest in the figure of the earth spread throughout Europe. Later Newton established his theorem showing the earth to be an oblate spheroid. Three determinations of the earth's figure completed under the auspices of the French Academy of Science at the end of the eighteenth century proved Newton's concept correct. The new era of geodesy, the science of the figure of the earth, began. Ideas and definitions were developed and international organizations of geodesists were formed where controversial items could be discussed, and ideas and information exchanged.

The geodetic question of the dimensions and shape of the earth has always been of basic scientific interest. However, it also has an increased significance for modern air traffic, for rocketry, and in other fields. Particularly

for ballistic missiles, where high altitudes and long distances are involved, the exact knowledge of position as well as accuracy of vertical and horizontal directions have a decisive importance. More accurate knowledge of the earth's surface is becoming a necessity.

The surface of the earth consists of lands and seas, and it is customary to call the surface formed by mean sea level the geoid. The geoid is thus an approximation to the earth's shape and size; the real earth's surface is irregular with Rocky Mountains above the level geoid, and Death Valley below. Although the geoid has a rather abstract definition, nevertheless it has physical reality. (See figure 1.2)

The geoid is a surface that is difficult to represent mathematically and therefore cannot conveniently be used as a reference surface for navigation, surveying and mapping. In its place a surface capable of mathematical representation is adopted, with a shape which closely approximates the geoid. Such a surface is a spheroid of reference whose geometrical centre coincides with the centre of gravity of the geoid and whose north-south axis is identical with the earth's axis of rotation.

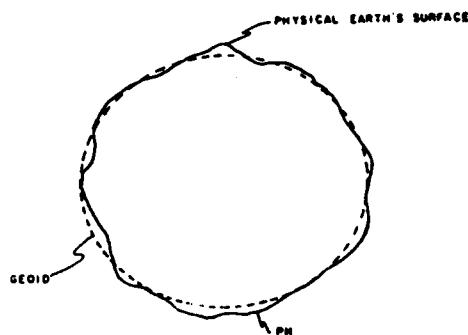


Fig. 1.2-The geoid.

In the twentieth century the basic Eratosthenesian idea of arc measurement is still followed in determining the figure of the earth. The direct measurement of a distance is replaced by a chain of triangles as represented in figure 1.3, or derived from a geodetic network of an area. Of course, astronomical observations are included. The arc is recalculated as though all points of the arc were transferred to the geoid at sea level. (See figure 1.3)

Several different arc determinations produce various radii of the earth's curvature, and it is through the mathematical reconciliation of the arcs

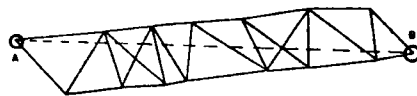


Fig. 1.3-Geodetic arc measurement between A and B.

with the corresponding radii of curvature that we have come to adopt as the reference figure the so-called spheroid of reference. It can be expected that the resulting spheroid will be the "best fit" only to the survey data used to determine it, and will always be an approximation valid only in the area of the surveys. But the spheroid of reference has great advantages over the geoid: maps can be shown on it; angles and distances can be computed; and the deviations of the geoid from the spheroid can be determined with certain accuracy.

In Table 1.1 the dimensions of some reference spheroids deduced from geodetic surveys, are given.

TABLE 1.1

Dimensions of Earth's Spheroids

Determination		Equatorial Semi Axis metres	Polar Flattening
Bessel	1841	6,377,397	1:299.15
Clarke	1866	6,378,206	1:294.98
Hayford	1909	6,378,388	1:297.0
Krassowski	1942	6,378,295	1:298.4
Hough	1956	6,378,270	1:297
Krassowski (U.S.S.R. reference spheroid)		6,378,245	1:298.3

In North American surveys, the reference figure used is that of Clarke, computed in 1866. This spheroid of 1866 differs somewhat from others, but in the area of North America it offers a fairly good approximation to the geoid.

As mentioned previously the spheroid of reference is only an approximation and there exist deviations of the geoid with respect to the spheroid which can be expressed in terms of the deflections of the geoid-vertical and in elevations of the geoid above the spheroid. The deflection of the vertical is the angle enclosed between the plumb bob hanging at the station and the perpendicular to the spheroid at this station. If astronomical data is compared with the geodetic data, these discrepancies can be expressed in terms of latitude and longitude, azimuth and distance. Fortunately there is a method for comparing the astronomical and geodetic data in order to control the precision of surveys. The deviation of the plumb line is caused by topography in the area surrounding the station and the variation of density in the earth's crust.

The difference between the apparent motion of stars and the period of the moon's revolution is the basis for lunar methods in geodesy. Information about the earth is provided by calculating the position of the moon among the reference stars of the celestial sphere. The study of the moon in eclipse occurring as the earth's shadow passes over it also supplies data for location

evaluation of the earth. Both of these methods suffer, however, because of imprecise knowledge of the moon's profile and of the distance between the centres of gravity of the moon and the earth. The same can be said about the sun's eclipses. The lunar methods have many practical advantages and have been largely applied in the past. Lately there has been a revival of interest in lunar methods. It has been proposed to use a lunar camera to photograph the moon against the stellar background. Then from known stars the position of the moon in terms of local co-ordinates of the station can be computed.

The arc measurement, position determination from star observation, and lunar methods may be considered as mainly geometric approaches to the determination of the figure of the earth. The physical approach to the determination of the shape of the earth is based primarily on the measurement of gravity variations on the surface of the earth.

The gravity value measured at a station, by means of a pendulum or another kind of a gravimeter, depends mainly on the geographic position of the station and on attraction of local and distant masses. The comparison of gravity data deduced from the assumption that the earth is a perfect spheroid, and the value of gravity at sea level, enables us to gather information about the actual shape of the earth. A great amount of gravity information has been collected already by geologists. In addition, ingenious devices for use on ships and in submarines make gravity observations possible on the sea, whereas other geodetic surveys and astronomical observations are limited of course to land areas. It is possible to detect plumb deviations from the gravity anomalies with a precision of a fraction of a second of arc.

Besides the question of the figure and shape of the earth, its gravity centre, etc., there exists the geodetic problem of determining the relative positions of continents. When we undertake a survey for map making purposes involving the stretching of long chains of triangles across the continent,

we rely upon measurements of distances and of directions with extremely precise surveying instruments. But the measurements have been made only as far as the shoreline of the continents. There they stop.

As indicated before, a spheroid of reference is calculated to represent closely the size and shape of the earth in the survey area. Map makers of different countries have used surfaces of different shapes and dimensions.

The problem arises of how to connect one set of geodetic latitude and longitude data across the unmarked ocean to another set of geodetic values related to different spheroids of reference and having different Datum systems--a vital question for those concerned with aiming ballistic missiles.

What we need in this instance is a network of some kind to cross the oceans. This would require measurements of either distances or angles. A radar reflector or a slave transmitter placed on the moon's surface would

TABLE 1.2

Distance km.	Altitude km.
1000	20
2000	80
3000	181
4000	328
5000	525
6000	779
7000	1101

provide a good apex for intercontinental ties enabling the geodesist to measure accurately the distance from various places on both coasts of the ocean to the moon.

The advent of the artificial satellite has already opened up new possibilities to the geodesist. He sees in it properties which are useful in triangulating a connection between two geodetic surveys. The satellite, due to its altitude above the ground, can be observed simultaneously from two widely separated points.

Table 1.2 gives the altitudes required for observations of the satellite seen at the same time in the horizon of two stations. The approximate data is based on the assumption that the earth's radius is 6,370 km.

In respect to the above table it may be noted that the shortest distance

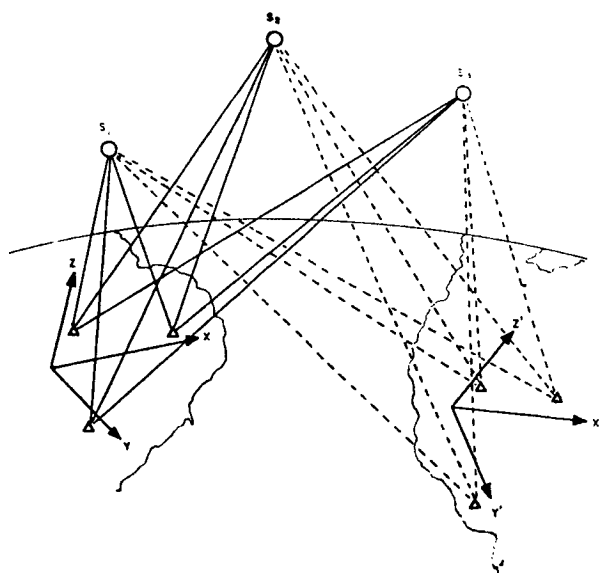


Fig. 1.4-Intercontinental geodetic ties.

between South America and Africa is around 3,000 km., where 1 km. is equivalent to 0.621 mile.

Suppose then that we make simultaneous distance measurements from a group of stations on each coastal area upon the orbiting satellite. From a number of simultaneous and non co-linear observations the separate geodetic systems of reference may be correlated. (See figure 1.4)

The difficulties of simultaneous observations while tracking a body travelling at great speed are considerable. However, the artificial satellite programme offers a great opportunity in the solution of earth survey problems. Once a synchronized observation system is arranged, satellites will provide a continuing opportunity for observations. Correlation of the maps of separate

continents will solve a problem that has eluded practical geography since its beginning.

The scientific significance of satellites for geodetic research has been recognized by the U.S. National Academy of Sciences. As a result the Geodesy Committee of Space Science Board has been formed. In the committee's opinion the following geodetic objectives could be advanced through satellite observations:

- (a) Correlation of position of widely separated geodetic datums. An accuracy of 30 metres is desired and believed possible to obtain.
- (b) An evaluation of the size and shape of the earth.
- (c) The determination of the gravity field of the earth.
- (d) The position of the gravitational centre of the earth.

Presently, extensive studies related to geodetic satellites are being carried out in the United States. They are indications that a special satellite for geodetic applications will become a reality in the not-too-distant future.

It may be mentioned that non-geodetic satellites have already yielded valuable information about the flattening of the earth as being 298.24. This figure is in exact agreement with Krassowski's flattening number (see Table 1.1). Recent orbital analysis of Vanguard, 1958 **B**2 indicates that the earth has a pear shape, with 15 metres of undulation in the geoid.

The rotational ellipsoid of reference is defined by the following parameters:

- orientation of the ellipsoidal axis with respect to the axis of rotation of the Earth; this requirement is actually identical with the angle between the ellipsoidal and terrestrial equators. This angle is usually defined in terms of two component angles.

- two ellipsoidal parameters, usually the semi-major axis, a , and either the eccentricity, e , or the flattening, f .

- position of the Earth's mass center with respect to the geometrical center of the reference ellipsoid (three parameters).

Therefore, seven parameters are required to relate the rotational ellipsoid of reference to the Earth's figure. It is quite apparent that by abandoning the approximation that the Earth is a rotational figure (e. g. assuming the equator is elliptical) and introducing additional parameters for the reference surface, a better approximation to the real Earth can be obtained. As is customary in physical sciences, we can either attempt to improve the approximation by increasing the complexity of the mathematical model; or we can define the discrepancies between the physical Earth and an accepted but somewhat simplified mathematical model. In geodesy, for the sake of computational advantages, the latter approach is taken.

In geometrical geodesy, where geometrical relationships are involved, it is sufficient to approximate the Earth's figure with a rotational ellipsoid. However, in dynamic geodesy, where the concern is with forces and accelerations, the ellipsoidal model of the Earth is not satisfactory. In this case, the Earth is approximated with an n-th order spheroid.

In this chapter, we are concerned with the geometrical considerations of relating points on the surface of the ellipsoid, while assuming that the relative positions of the points on the Earth remain unchanged.

2. Meridional Ellipse

The equation of the ellipsoid of revolution is usually written in rectangular coordinates as follows,

$$\frac{x^2+y^2}{a^2} + \frac{z^2}{b^2} = 1 \quad (2.1)$$

The origin of coordinates is located in the center of the ellipsoid. An arbitrary plane intersects this ellipsoid along circles and ellipses. A plane perpendicular to the equatorial plane and containing the minor axis of the ellipsoid is called the meridional plane; it intersects the ellipsoid along the meridional ellipse.

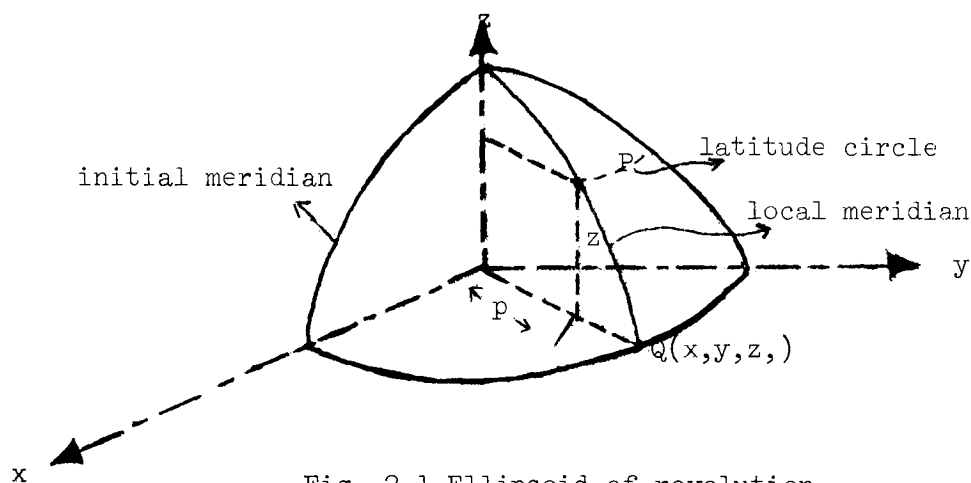


Fig. 2.1-Ellipsoid of revolution

Denoting the rectangular coordinates of a point, P, of the meridional ellipse by p, and z, the equation of the meridional section can be written as that of an ellipse:

$$\frac{p^2}{a^2} + \frac{z^2}{b^2} = 1 \quad (2.2)$$

where

$$p^2 = x^2 + y^2$$

3. Reduced Latitude

The reduced latitude, β , is an angle denoted as follows,

$$\frac{z}{b} = \sin \beta \quad \frac{p}{a} = \cos \beta$$

The reduced latitude is a parameter used in the construction of an ellipse given by its semi-major axis, a , and semi-minor axis, b . The procedure is indicated in figure 3.1.

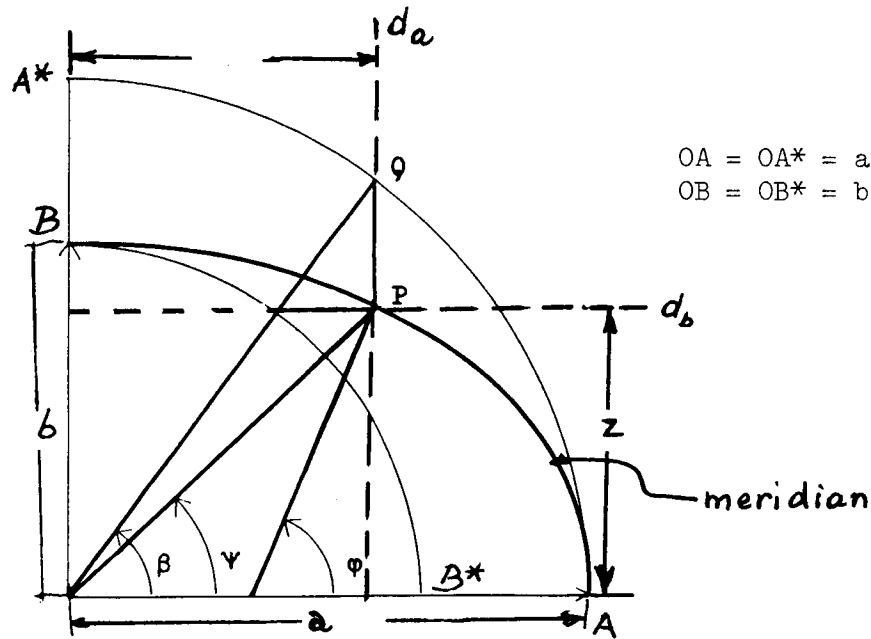


Fig. 3.1-Definition of reduced latitude β

Two concentric circles have their centers in O, their respective radii are a and b . The ellipse point P is formed by the intersection of perpendiculars d_a and d_b . The rest of the construction is self-explanatory. By inspecting figure 3.1 we find that

$$\begin{aligned} p &= a \cos \beta \\ z &= b \sin \beta \end{aligned} \tag{3.1}$$

$$\frac{z}{p} = \frac{b}{a} \tan \beta$$

It is obvious that β satisfies the relation

$$\sin^2 \beta + \cos^2 \beta = 1$$

4. Geodetic (Geographic) Latitude

The geodetic latitude, φ , of a point, P, is defined as the angle between the equatorial plane and the normal to the ellipsoid at the point, P. Figure 4.1 illustrates a meridional section of the ellipsoid. The tangent to the meridional ellipse, t, which passes through the point, P, encloses an angle of $90^\circ + \varphi$ with the positive direction of the p-axis.

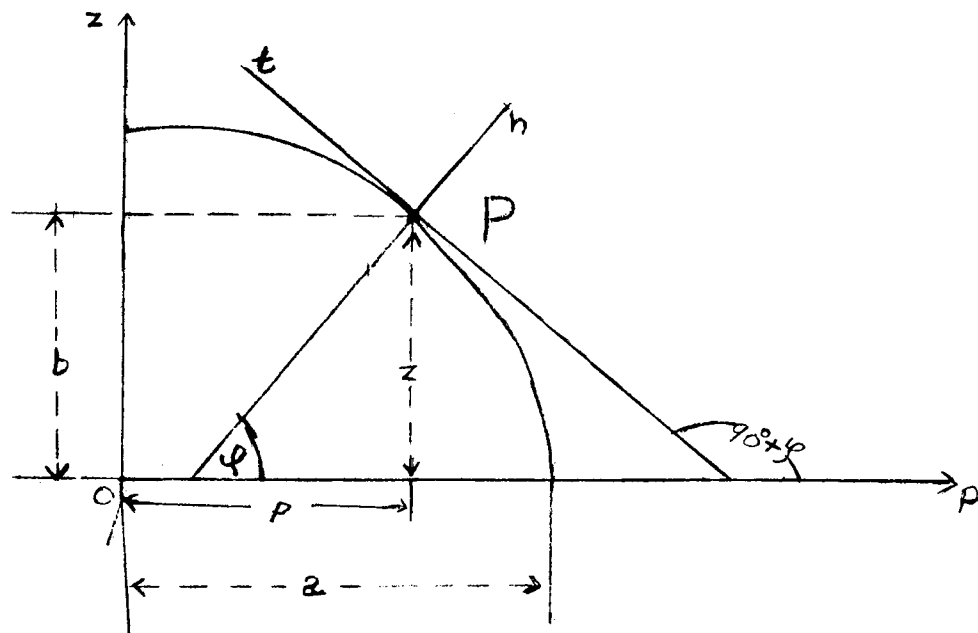


Fig. 4.1-Geodetic latitude

The differentiation of the equation of the meridional ellipse (eq. 2.2) with respect to the parameters p and z yields

$$\frac{dz}{dp} = -\frac{b^2 p}{a^2 z} = -\cot \varphi \quad (4.1)$$

By expressing $\cot \varphi$ in terms of sines and cosines, the following equation for the meridional ellipse in terms of geodetic latitude φ is obtained:

$$p^2 b^4 \sin^2 \varphi + z^2 a^4 \cos^2 \varphi = 0$$

Comparing the above equation with that for the meridional ellipse in terms of the parameters, p and z ($b^2 p^2 + a^2 z^2 = a^2 b^2$), the geocentric rectangular coordinates of an ellipsoid point are obtained,

$$p = \frac{a^2 \cos \varphi}{\sqrt{a^2 \cos^2 \varphi + b^2 \sin^2 \varphi}} ; \quad z = \frac{b^2 \sin \varphi}{\sqrt{a^2 \cos^2 \varphi + b^2 \sin^2 \varphi}} \quad (4.2)$$

The above equation is the parametric representation of the ellipsoid in terms of geodetic latitude, φ . Remembering that

$$p^2 = x^2 + y^2$$

$$b^2 = a^2(1 - e^2)$$

and introducing the denotation

$$W^2 = 1 - e^2 \sin^2 \varphi \quad (4.3)$$

the coordinates, p and z , can be written as

$$p = \frac{a \cos \varphi}{W} ; \quad z = \frac{a(1 - e^2) \sin \varphi}{W} = \frac{b \sqrt{1 - e^2} \sin \varphi}{W} \quad (4.4)$$

5. Relations Between the Reduced and Geodetic Latitudes

The comparison of equations (3.1) and (4.2) yields

$$\frac{a}{b} \frac{z}{p} = \frac{b}{a} \tan \varphi = \tan \beta$$

or introducing the expression of the eccentricity,

$$e^2 = \frac{a^2 - b^2}{a^2}$$

the following expression is obtained:

$$\tan \beta = \sqrt{1 - e^2} \tan \varphi \quad (5.1a)$$

From trigonometric relations, the following equations can be derived:

$$\cos \beta = \frac{1}{W} \cos \varphi \quad (5.1b)$$

$$\sin \beta = \frac{\sqrt{1 - e^2}}{W} \sin \varphi \quad (5.1c)$$

The differentiation of both sides of equation (5.1a) with regard to β and φ respectively leads to the relation

$$\frac{d\varphi}{d\beta} = \frac{W^2}{\sqrt{1 - e^2}} \quad (5.2)$$

In order to derive the difference between the reduced latitude and the geocentric latitude we write the trigonometric relation

$$\sin (\varphi - \beta) = \sin \varphi \cos \beta - \cos \varphi \sin \beta$$

Inserting the relations (5.1b) and (5.1c), there results the expression

$$\sin (\varphi - \beta) = \frac{f \sin 2\varphi}{2N} \quad (5.3)$$

where f is the ellipsoidal flattening and is defined as

$$f = \frac{a - b}{a} \quad (5.4)$$

or, in terms of the eccentricity, e , it is defined by

$$f = 1 - \sqrt{1 - e^2} \quad (5.4a)$$

It must be noted that for the terrestrial ellipsoid of reference $\varphi - \beta$ is always a positive quantity and smaller than 350 seconds of arc. In the first approximation,

$$\varphi - \beta \approx \frac{e^2}{4} \sin 2\varphi \quad (5.5)$$

in radians. In order to obtain the difference of latitudes in seconds of arc, multiply the right side of equation (5.5) by 206265.

6. Geocentric Latitude

The geocentric latitude, ψ , of a point, P , is defined as the angle between the equatorial plane and the radius vector of the point, P , on the ellipsoid. See figure 6.1.

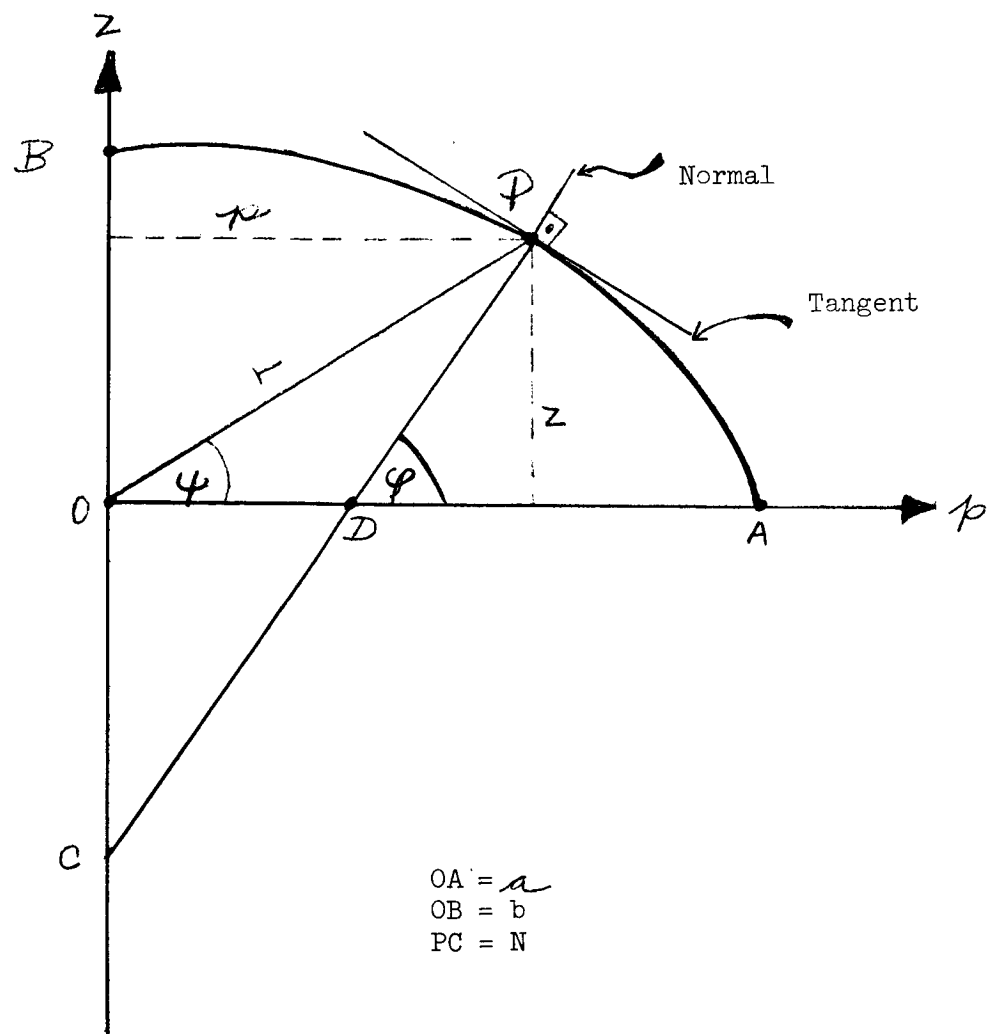


Fig. 6.1-Geocentric and geodetic latitudes

Interrelating the coordinates of a point, P , on the meridional ellipse, (p and z), the geocentric latitude (ψ), and the radius vector (r), leads to the following equations:

$$\begin{aligned} p &= r \cos \psi \\ z &= r \sin \psi \end{aligned} \tag{6.1}$$

$$\tan \psi = \frac{p}{z} \quad (6.2)$$

Utilizing the trigonometric identity,

$$\tan (A - B) = \frac{\tan A - \tan B}{1 + \tan A \tan B}$$

and the equation,

$$\tan(\psi) = (1 - e^2) \tan \varphi$$

(compare equations 1.6 and 1.12), there results the expression

$$\tan (\varphi - \psi) = \frac{e^2 \tan \varphi}{1 + (1 - e^2) \tan^2 \varphi} \quad (6.3)$$

Expanding the tangent on the right side of (6.3) and taking into consideration the fact that $\varphi - \psi$ is a small quantity, the following relation is satisfactory for most applications:

$$\varphi - \psi = \frac{e^2}{2} \sin 2\varphi \quad (6.4a)$$

or, equivalently,

$$\varphi - \psi = f \sin 2\varphi \quad (6.4b)$$

Both quantities are given in radians, in order to convert them into seconds of arc, the right sides of the above equations must be multiplied by 206265. An inspection of equations (6.4) indicates that, for $\varphi = 45^\circ$, $(\varphi - \psi)$ reaches a maximum of about 700 seconds of arc.

The radius vector, r , can be computed from the relation

$$r^2 = p^2 + z^2$$

Substituting for p and z from equation (4.4), the following expression is obtained for the radius vector of a point on the ellipsoid:

$$r^2 = \left(\frac{a}{w}\right)^2 [1 + e^2(e^2 - 2) \sin^2 \varphi] \quad (6.5)$$

7. Principal Radii of Curvature

The principal radii of curvature of the ellipsoid are located in the meridional (N-S direction) and in the prime vertical (E-W direction) planes of the point.

The meridional section of the ellipsoid is an ellipse. Defining the equation of the meridional ellipse as $z = F(p)$

its radius of curvature at the point P is given by the formula

$$M = - \frac{[1 + (z')^2]^{3/2}}{z''}$$

Since $z' = -\cot \varphi$ and $z'' = \csc^2 \varphi$ it follows that

$$M = \frac{a(1 - e^2)}{w^3} \quad (7.1)$$

The intersection of the prime vertical plane with the ellipsoid is also an ellipse, which is perpendicular to the meridional ellipse and tangential to the latitudinal circle of the point P . The radius of curvature of the

prime vertical ellipse at the point P is

$$N = p \sec \varphi = \frac{a}{W} \quad (7.2)$$

According to Euler's theorem, the radius of curvature in an arbitrary direction is given by the relationship

$$\frac{1}{R_A} = \frac{\cos^2 A}{M} + \frac{\sin^2 A}{N} \quad (7.3)$$

where A represents the angle between the North and the arbitrary direction.

8. Mean Spherical Radius of the Earth

In many problems it is convenient to represent the Earth as a sphere. The following mean terrestrial radii are feasible.

(a) The radius of a sphere whose surface area is equal to the area of the reference spheroid. This radius can be obtained by comparing the area of the sphere with that of the ellipsoid; the following equation can be written:

$$r_s^2 = a^2 \left(1 - \frac{e^2}{3} - \frac{e^4}{15} - \frac{e^6}{35} - \frac{e^8}{63} - \dots \right)$$

with the result that

$$r_s = a \sqrt{1 - \frac{e^2}{3} - \frac{e^4}{15} - \frac{e^6}{35} - \frac{e^8}{63} - \dots} \quad (8.1)$$

is the radius of the sphere in question.

(b) The radius defined as the average of two equal semi-major axis and the semi-minor axis of the spheroid. This radius is then defined as

$$r_m = \frac{2a + b}{3}$$

or, equivalently,

$$r_m = \frac{a}{3} (3 - f) \quad (8.2)$$

(c) The radius of a sphere whose volume is identical with that of the reference spheroid. It can be derived by equating the volume formulas for a circle and the spheroid:

$$r_v^3 = a^2 b$$

with the result that

$$r_v = a \sqrt[6]{1 - e^2} \quad (8.3)$$

By inserting the values for ellipsoidal parameters into each of the radius-defining equations, the respective spherical radii are obtained. The average earth's radius is:

$$R = 6371 \text{ km.}$$

9. Parametric Representation of the Ellipsoid in Terms of Geodetic Latitude and Longitude

The geodetic latitude has been defined in Section 3. The geodetic longitude is defined as the angle between the plane of the local meridian and the plane of the meridian of Greenwich.

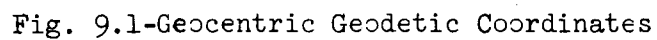
The rectangular geocentric coordinates (x,y,z) of a point, P, on the ellipsoid can be expressed in terms of geodetic latitude, longitude:

$$\begin{aligned}x &= N \cos \varphi \cos L \\y &= N \cos \varphi \sin L \\z &= N(1-e^2)\sin\varphi\end{aligned}\tag{9.1}$$

where, $N = CP$, is the radius of the curvature in the vertical, local East-West direction (See figure 1.8):

$$N = \frac{a}{\sqrt{1-e^2\sin^2\varphi}}$$

The system of coordinates is illustrated in figure (9.1) on the following page.



II

THE EARTH'S EXTERNAL GRAVIPOTENTIAL FIELD

The gravitational force is determined by the mass distribution of the attracting bodies. According to Newton's law of universal gravitation, the attraction is directly proportional to the product of the masses and inversely proportional to the square of the distance between them. The theory of the Newtonian potential approaches the problem by defining a gravitational field produced by the distribution of mass of the bodies. Notice that, in this definition, the geometrical shape of the body is not included; the definition is concerned with the mass distribution interior to and on the surface of the body.

The force, F , with which a mass point $Q(a,b,c)$ with mass, m , attracts an arbitrary mass point $P(x,y,z)$ of mass, m^* , is defined as

$$F = G \frac{m m^*}{r^2}$$

where G is the universal gravitational constant and $r^2 = (x-a)^2 + (y-b)^2 + (z-c)^2$. The dimensions of the gravitational constant, G , are $\text{gm}^{-1}\text{cm}^3\text{sec}^{-2}$. The gravitational constant is determined by measuring the force, F , for known masses, m and m^* , and the distance, r , between them. Two such experiments were those of Cavendish (1798), who measured the attraction of two spheres located in the same horizontal plane, and those of Joly (1881), who placed one sphere above the other. The gravitational constant, as determined by the latest experiments, is equal to $0.6685 \text{ gm}^{-1}\text{cm}^3\text{sec}^{-2}$.

10. Properties of Potentials

The gravipotential is a function of the coordinates in the three dimensional space and can be represented symbolically by $V(x,y,z)$. The partial derivatives of the potential with respect to the rectangular coordinates equal the vectorial components of the force vector. The force vector can be written in the form

$$\vec{F}(x,y,z) = \frac{\partial V}{\partial x} \vec{i} + \frac{\partial V}{\partial y} \vec{j} + \frac{\partial V}{\partial z} \vec{k} \quad (10.1)$$

The potential has dimensions of $\text{gm cm}^2 \text{ sec}^{-2}$. The dimensions for the potential corresponding to the vector acceleration are $\text{cm}^2 \text{ sec}^{-2}$; the acceleration is equal to force per unit mass, and the corresponding potential is equal to the gravipotential per unit mass.

The total differential of $V(x,y,z)$ is given by the relation

$$dV = \frac{\partial V}{\partial x} dx + \frac{\partial V}{\partial y} dy + \frac{\partial V}{\partial z} dz$$

By denoting the cosines of the angles between the direction, \vec{L} , and the coordinate axis (direction cosines) as

$$\frac{dx}{dL} = \cos(x,L) \quad \frac{dy}{dL} = \cos(y,L) \quad \frac{dz}{dL} = \cos(z,L)$$

the derivative of the potential with respect to the direction, \vec{L} , can be written as

$$\frac{dV}{dL} = \frac{\partial V}{\partial x} \cos(x,L) + \frac{\partial V}{\partial y} \cos(y,L) + \frac{\partial V}{\partial z} \cos(z,L) \quad (10.2)$$

By reason of equation (10.1)

$$\frac{\partial V}{\partial x} = F \cos(\bar{F}, x) \quad \frac{\partial V}{\partial y} = F \cos(\bar{F}, y) \quad \frac{\partial V}{\partial z} = F \cos(\bar{F}, z)$$

and therefore the total differential of the potential in the arbitrary direction, \bar{L} , can be written as

$$dV = F \cos(\bar{F}, dL) dL \quad (10.3)$$

From this equation the following properties of the potential can be deduced:

(1) The derivative of the potential in a direction is equal to the projection of the vector force (or acceleration) on that direction

$$\frac{dV}{dL} = F \cos(\bar{F}, dL) \quad (10.4)$$

(2) The work that the force, \bar{F} , does in displacing a unit mass ($m^*=1$) from a point, P_1 , to another point, P_2 , is equal to the difference of the potentials,

$$V_1 - V_2 = \int_{(1,2)} F \cos(\bar{F}, dL) dL \quad (10.5)$$

(3) Inspection of equation (10.4) yields the conclusion that the differential of the potential can be either positive or negative. The sign of the potential increment depends upon the direction of the displacement, dL , relative to the vector, \bar{F} . The maximum of the potential increment occurs when $\cos(F, dL) = +1$; the minimum occurs when $\cos(F, dL) = -1$; and the displacement is perpendicular to \bar{F} , or equivalently, when $\cos(F, dL) = 0$, the increment of the potential, dV , equals zero; this also occurs when $V(x, y, z)$ is a constant along the displacement, or $V(x, y, z) = \text{constant}$.

Surfaces on which the potential is constant are called equipotential surfaces; these surfaces have the property that the vector force at a point is perpendicular to the equipotential surface intersecting the point. For a single valued potential the equipotential surfaces do not interject; for a continuous potential the equipotential surfaces are closed or be continued to the boundary of the existence of the potential.

(4) The potential decreases most rapidly in the direction for which $\cos(\mathbf{F}, d\mathbf{L}) = -1$. Denoting the corresponding displacement by dh , it follows from equation (10.3) that

$$dh = -\frac{dV}{F} \quad (10.6)$$

It is obvious that in general the force \bar{F} on the equipotential surface varies, hence the separation between two equipotential surfaces is not constant. The vector force lies along the displacement dh ; and equation (10.6) can be written in the form

$$dV = -Fdh \quad (10.6a)$$

Integration of this equation between two equipotential surfaces results in the equation

$$C_2 - C_1 = - \int_{h_1}^{h_2} F dh$$

Taking advantage of the fact that the sign of dh does not change, the following equation can be written,

$$C_2 - C_1 = - F_m dh$$

where F_m is the average value of the force, F , along the line of force.

11. The Force of Gravity

Let us consider a solid body which is composed of molecules with various densities; let us also use a rectangular coordinate system whose origin is at the center of mass and whose axis coincide with the principal axis for the moments of inertia. We then define the volume element at a point $P'(x',y',z')$ as,

$$dT = dx' dy' dz'$$

and the density as

$$\theta = \theta(x', y', z')$$

with the result that an element of mass is

$$dm = \theta dx' dy' dz'$$

By reason of the choice of coordinate systems, the static moments for the entire body vanish:

$$\sum x' dm = \sum y' dm = \sum z' dm = 0$$

or, in integral form:

$$\begin{aligned} \int x' \theta dx' dy' dz' &= 0 \\ \int y' \theta dx' dy' dz' &= 0 \\ \int z' \theta dx' dy' dz' &= 0 \end{aligned} \tag{11.1}$$

Also, the deviation (centrifugal) moments with respect to the chosen coordinate

system vanish:

$$\begin{aligned} \mathcal{D}_{yz} &= \int_T y' z' \Theta dx' dy' dz' = 0 \\ \mathcal{D}_{zx} &= \int_T z' x' \Theta dx' dy' dz' = 0 \\ \mathcal{D}_{xy} &= \int_T x' y' \Theta dx' dy' dz' = 0 \end{aligned} \quad (11.2)$$

The principal moments of inertia can be written as

$$\begin{aligned} A &= \int_T (y'^2 + z'^2) \Theta dx' dy' dz' \\ B &= \int_T (z'^2 + x'^2) \Theta dx' dy' dz' \\ C &= \int_T (x'^2 + y'^2) \Theta dx' dy' dz' \end{aligned} \quad (11.3)$$

And the total mass can be written in the form

$$M = \int_T \Theta dx' dy' dz' \quad (11.4)$$

and the external gravity potential due to the body T can be expressed as

$$V(x, y, z) = G \int_T \frac{\Theta dx' dy' dz'}{r} = G \int_T \frac{\Theta(x', y', z') dx' dy' dz'}{[(x-x')^2 + (y-y')^2 + (z-z')^2]^{1/2}} \quad (11.5)$$

and the potential has the properties that it and its first derivative are finite and continuous; this implies that the second derivatives exist.

The potential has the added property that it vanishes at infinity.

$$V(\infty) = 0$$

Let an external point have the coordinates (x, y, z) . Then the partial derivatives of the potential are:

$$V_x = \frac{\partial V}{\partial x} = G \int_T \frac{\theta (x' - x) dx' dy' dz'}{r^3}$$

$$V_y = \frac{\partial V}{\partial y} = G \int_T \frac{\theta (y' - y) dx' dy' dz'}{r^3}$$

$$V_z = \frac{\partial V}{\partial z} = G \int_T \frac{\theta (z' - z) dx' dy' dz'}{r^3}$$

where

$$r^2 = (x' - x)^2 + (y' - y)^2 + (z' - z)^2 \quad (11.6)$$

These expressions define the three components of the gravitational force in the direction of coordinate axis.

Let us now assume that the body rotates around the principal axis for which the moment of inertia is the greatest (call it the z-axis). A unit mass located at the external point $P(x, y, z)$ will undergo a centrifugal force

$$F_Q = \omega^2 \sqrt{x^2 + y^2}$$

This force can be obtained by taking the gradient of the following scalar potential:

$$Q = \frac{1}{2} \omega^2 (x^2 + y^2) \quad (11.7)$$

Then, the total force acting on the unit mass can be obtained from the scalar sum of the two potentials,

$$U = V + Q$$

or, equivalently,

$$U = G \int_T \frac{\theta dx' dy' dz'}{r} + \frac{1}{2} \omega^2 (x^2 + y^2) \quad (11.8)$$

The corresponding T acceleration of the unit mass equals the force and is

$$\vec{g} = \frac{\partial U}{\partial x} \vec{i} + \frac{\partial U}{\partial y} \vec{j} + \frac{\partial U}{\partial z} \vec{k} \quad (11.9)$$

The components of the acceleration vector are

$$\begin{aligned} g_x &= G \int_T \frac{\theta (x' - x) dx' dy' dz'}{r^3} + \omega^2 x \\ g_y &= G \int_T \frac{\theta (y' - y) dx' dy' dz'}{r^3} + \omega^2 y \\ g_z &= G \int_T \frac{\theta (z' - z) dx' dy' dz'}{r^3} \end{aligned} \quad (11.10)$$

and the Laplacian has the value

$$\nabla^2 U = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 2\omega^2 \quad (11.11)$$

In analogy to equation (10.6a), the increase in the potential due to a displacement of the unit mass by a distance dL is

$$dU = g \cos(\theta, dL) dL \quad (11.12)$$

and, if the displacement is along the normal to the equipotential at the point, the increase is

$$dU = -g dh \quad (11.13)$$

or, if an expression for the displacement is desired,

$$dh = -\frac{dU}{g} \quad (11.14)$$

The last equation is convenient for computing undulations, dh , of the equipotential surfaces ($U=\text{constant}$) which correspond to changes or perturbations, dU , of the potential U .

12. Spherical Harmonics Expansion of the Gravipotential

We would like to expand the potential

$$U = G \int \frac{dm}{r} + \frac{1}{2} \omega^2 (x^2 + y^2) \quad (12.1)$$

in spherical harmonic functions, which are the functions that frequently occur when functions in physics are expressed in spherical polar coordinates.

On the basis of the figure

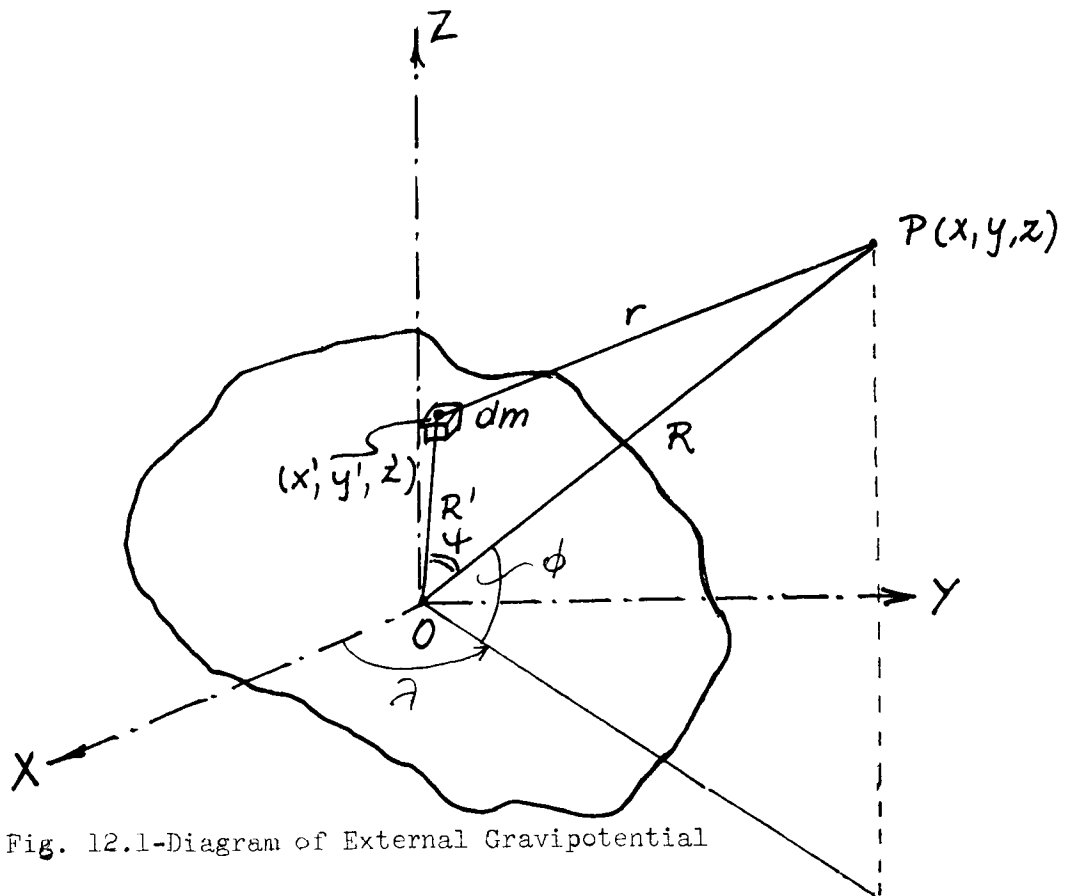


Fig. 12.1-Diagram of External Gravipotential

r^2 is defined by the equation

$$r^2 = R^2 + R'^2 - 2RR'\cos\phi$$

Then, the reciprocal of the radius,

$$\frac{1}{r} = (R^2 + R'^2 - 2RR'\cos\psi)^{-1/2}$$

can be written as

$$\frac{1}{r} = \frac{1}{R} \left[1 + \left(\frac{R'}{R}\right)^2 - 2\left(\frac{R'}{R}\right)\cos\psi \right]^{-1/2}$$

When $R \gg R'$, the above expression can be expanded in spherical harmonics to give

$$U = \frac{G}{R} \left[\int_T dm + \frac{1}{R} \int_T R' P_1(\cos\psi) dm + \frac{1}{R^2} \int_T R'^2 P_2(\cos\psi) dm + \dots + \frac{1}{R^n} \int_T R'^n P_n(\cos\psi) dm + \frac{1}{2} \omega^2 (x^2 + y^2) \right] \quad (12.2)$$

For $U = \text{constant}$, equation (12.2) represents an equipotential surface called the spheroid. The first integral, $\int_T dm = M$, represents the mass of the earth. The second term can be rewritten, by reason of

$$P_1(\cos\psi) = \cos\psi = \frac{xx' + yy' + zz'}{RR'}$$

to give

$$\int_T R' P_1(\cos\psi) dm = \frac{1}{R} \left[x \int_T x' dm + y \int_T y' dm + z \int_T z' dm \right]$$

Since the origin is at the center of mass,

$$\int_T x' dm = \int_T y' dm = \int_T z' dm = 0$$

and thus the second term vanishes:

$$\int_T R' P_1(\cos\psi) dm = 0$$

Using trigonometric transformations, the third integral can be rewritten to give:

$$\begin{aligned} \int_T R_2' P_2(\cos \psi) dm &= \frac{1}{2} (3 \sin^2 \phi - 1) \int_T (z'^2 - \frac{x'^2 + y'^2}{2}) dm \\ &+ 3 \sin \phi \cos \phi \left[\cos \lambda \int_T x' z' dm + \sin \lambda \int_T y' z' dm \right. \\ &\left. + \frac{3}{4} \cos^2 \phi \left[\cos 2\lambda \int_T (x'^2 - y'^2) dm + \sin 2\lambda \int_T 2x'y' dm \right] \right] \end{aligned}$$

By reason of the fact, and only by reason of the fact, that the coordinate axes are the principal axes of the moments of inertia, all terms except the first vanish. Using the cartesian-polar transformation equations,

$$x = R \cos \phi \cos \lambda$$

$$y = R \cos \phi \sin \lambda$$

$$z = R \sin \phi$$

$$x' = R' \cos \phi' \cos \lambda'$$

$$y' = R' \cos \phi' \sin \lambda'$$

$$z' = R' \sin \phi'$$

and eliminating the vanishing terms, the potential takes the form

$$U = \frac{GM}{R} + \frac{G}{2R^3} \left(C - \frac{A+B}{2} \right) (1 - 3\sin^2 \phi) + \frac{3G}{4R^3} (B-A) \cos^2 \phi \cos 2\lambda \\ + \frac{G}{R} \sum_{n=3}^{\infty} \left(\frac{R'}{R} \right)^n P_n(\cos \phi) dm + \frac{1}{2} \omega^2 R^2 \cos^2 \phi \quad (12.3)$$

The first term of the equation is called the spherical term and gives the potential of the earth were it perfectly spherical; the second term is related to the polar flattening; and the third is interpreted as the effect of equatorial ellipticity. The second term is a function of the latitude only, Such term is called zonal term; the third is a function of both the latitude and longitude; such a term is called tesseral term.

It is customary to write equation (12.3) in its transformed form,

$$U(r, \phi, \lambda) = \frac{GM}{r} \left[1 + \sum_{n=2}^{\infty} \sum_{m=0}^n \left(\frac{a}{r} \right)^n P_{nm}(\sin \phi) \right]_{nm} \cos m(\lambda - \lambda_{nm}) \quad (12.4)$$

where

r, ϕ, λ = the geocentric radius, latitude, and longitude of the external point

a = the mean equatorial radius = 6378166 meters

GM = the earth's gravitational parameter = $3.986075 \times 10^{14} \text{ m}^3 \text{sec}^{-2}$

P_n = the Legendre polynomial of degree n

P_{nm} = the associated Legendre function of the first kind of degree n and order m

λ_{nm} = the longitude associated with J_{nm}

J_n = the numerical zonal coefficients

J_{nm} = the numerical tesseral coefficients

$J_{10}=J_{11}=J_{21} = 0$ due to the assumptions made above about the choice of the earth's geocentric coordinates.

Leaving these terms out in computations introduces errors commensurate with the errors in our estimates of the location of the center of mass of the earth. If this center of coordinates is not near the very center, the harmonic expansion will diverge. This divergency causes difficulties in the application of spherical harmonic expansion for near earth gravity field.

The following values for the coefficients were obtained by scientists of the Applied Physics Laboratory of Johns Hopkins University by an analysis of satellite tracking data.

Table 12.1 J - Coefficients

n	m	J_{nm}	λ_{nm} (degrees) from Greenwich
2	0**	1082.2×10^{-6}	
	2*	1.72	-13.4
3	0**	2.645	
	1*	2.01	6.7
	2*	0.477	-14.6
	3*	0.165	18.7
4	0**	1.75	
	1*	0.679	-142.0
	2*	0.193	23.4
	3*	0.0506	0.2
	4*	0.006	34.5

*Source: C. A. Wagner, Journal of Geophysical Research, Vol. 71, No. 6, 1966, p. 1707.

**Source: Unpublished, Applied Physics Laboratory.

Table 12.1 shows that the term J_{20} is by far the largest. It also shows that the zonal coefficients are larger than the tesseral coefficients of the same order. Using different observations of different satellites, various scientists have produced values for the coefficients which not necessarily agree.

Table 12.1 contains coefficients of eleven terms of the harmonic expansion. It is quite apparent that eleven terms cannot provide accurate information about the fine structure of the gravity field due to geological, crustal, and topographical features of the earth. As it stands, it can be called a global gravity formula representing the gravity field due to the earth's interior mass distribution. A study was made by the author in which the earth was represented by a set of nonconcentric, nonoverlapping spherical shells. The only assumption made was that the centers of the spheres cluster around the origin of the coordinate system. Numerical calculations indicate that such a model gives results equivalent to those obtained from a spherical harmonic expansion, but without certain mathematical drawbacks of the spherical harmonic method (in particular, the problem of divergence of the series for near-earth points).

Satellites have proved to be a very useful tool in research of the gravipotential field; it is one of the major pay-offs of space research to date.

But like any scientific tool, the gravimetric satellite has limitations. It would be desirable to fly a low satellite in order to investigate regional gravity anomalies. However, errors associated with our ignorance of the perturbing effect of air drag at low altitudes contaminates the effect of gravity anomalies, with the result that reliable conclusions cannot be obtained. On the other hand, at high altitudes, where the air drag becomes insignificant, the details of the gravitational field are smoothed out since many local features contribute to the field. Therefore, great care must be exercised in using satellites for analysis of the near-earth gravity field, and for analysis of near-surface mass features.

One method of obtaining the fine structure of the gravity field would be to use a dense network of tracking stations in the region in question; but cost factors enter the picture. The local perturbation potential could be also obtained by the derivation of a perturbation formula (or tables of perturbations) for specific regions based on a mathematical analysis of the effect of crustal and topographical layers. The author has found that using this approach, the crustal and topographical features can be satisfactorily approximated by representing the features as simplified geometrical figures.

The uncertainties in the gravity model may persist for some time due to the fact that the anomalies are principally due to interior mass distributions, and the mathematical fact that an infinite number of mass distributions can produce the same gravity field. Nor can seismic studies offer a quick solution because seismograms are open to multiple interpretations.

If the problem of uncertainties in the gravitational model could not be circumvented, it would be critical where satellites are to be used for obtaining accurately the geodetic locations of selected ground stations. In such studies, the satellite is observed simultaneously by three or more ground stations using either optical or ranging methods. A procedure of simultaneous observation has the advantage of not requiring the knowledge of either the orbital parameters or the gravity field. Three satellite systems may be mentioned: SECOR, ECHO, and PAGEOS.

The radio ranging method is illustrated in figure 12.2. The basic geometrical figure in the computations is a tetrahedron. The satellite is observed by stations, A_i , B_i , C_i , at known locations and also by the new station R , whose position is to be determined. Three simultaneous ranges, a_i , b_i , c_i , enable the determination of the geodetic coordinates of the satellite, S_i . By making three such sets of observations, S_1 , S_2 , S_3 , three

ranges, r_1 , r_2 , and r_3 , and the location of the new station, R , can be determined. It is not necessary that the three sets of observations, S_1 , S_2 , and S_3 , be made on the same satellite pass; nor need the same three known stations be used on all sets of observations. Accuracies can obviously be improved by using more than three known ground stations for each set of observations, and also by using more than three sets of observations.

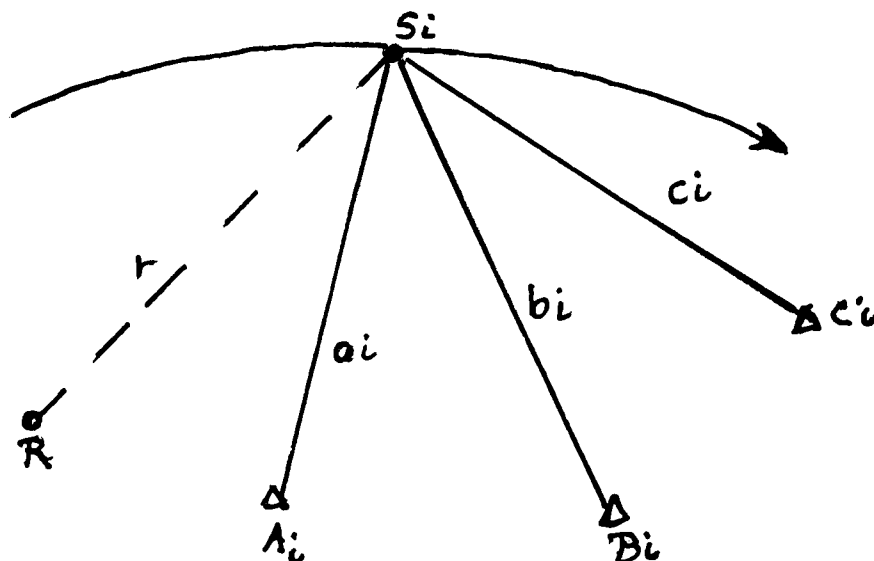


Fig. 12.2-Multiple-station radio tracking of a satellite

13. Gravitational Anomalies

The analytical function $U = U(r, \theta, \lambda)$, representing the potential has been given in Section 12. Due to regional mass distributions, the actual potential W_P at a point P deviates from the theoretical potential U .

The true potential W at a point P can be written as

$$W_P = U_P + T_P \quad (13.1)$$

where U_P and T_P are the respective theoretical and anomalous gravity potentials.

Constant potentials, $W = C$ and $U = C$ define respective equipotential surfaces. The surface corresponding to the theoretical potential $U = C$ is called a spheroid. A specific selected spheroid is called the reference spheroid; in geodesy, the spheroid of reference is a theoretical sea level, and the corresponding physical surface is defined as the geoid. It must be pointed out that this selection is a matter of convenience; it is possible to select a set of any other theoretical and physical equipotential surfaces such as $W = V = \text{constant}$.

The potential as such cannot be measured; the potential is expressed in terms of its gravity gradient. It is of interest to establish the relationship between the theoretical (spheroidal) gravity and the corresponding physical (true) gravity.

It is assumed that point Q on a spheroid corresponds to a point P on the true equipotential surface. Also it is assumed that the points P and Q are not too far apart (See figure 13.1). The respective accelerations due to potentials W, and U are

$$\bar{g} = \text{grad } W \quad \bar{\gamma} = \text{grad } U$$

Introduce the subscript convention that a potential with a subscript representing a point in space is to be evaluated at that point. Thus,

$$U_Q = U(Q)$$

We form the relation

$$\text{grad}_Q U = \text{grad}_P W = \gamma_Q g_P \cos E$$

where E is the angle between the vectors of spheroidal gravity γ_Q and actual gravity g_P . Angle E is therefore the local plumb deflection. Since

the points P and Q are close, the normal gravity at the point $P(x_p, y_p, z_p)$ can be represented in terms of a power series in the vicinity of the point $Q(x_a, y_a, z_a)$. Then

$$U_p = U_a + \frac{\partial U}{\partial x_a}(x_p - x_a) + \frac{\partial U}{\partial y_a}(y_p - y_a) + \frac{\partial U}{\partial z_a}(z_p - z_a) \quad (13.2)$$

The higher terms can be neglected because the points P and Q are close.

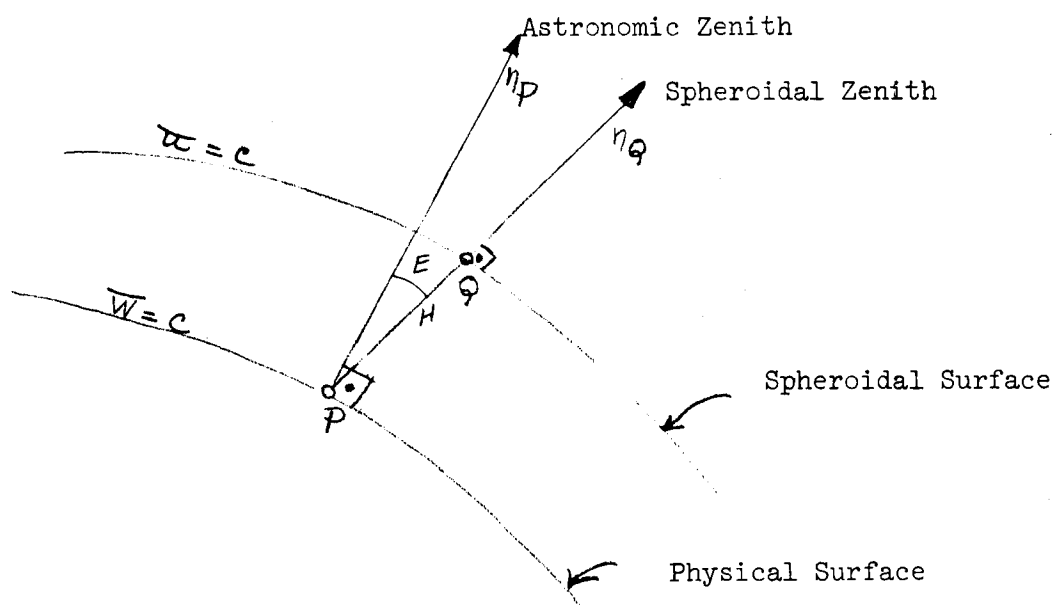


Fig. 13.1

Denoting by H the height of the true equipotential surface above the spheroid, and reckoning positive height in the upward zenith direction of the physical surface, we can write

$$\begin{aligned} x_p - x_a &= -\frac{H}{g} \frac{\partial W}{\partial x_p} \\ y_p - y_a &= -\frac{H}{g} \frac{\partial W}{\partial y_p} \\ z_p - z_a &= -\frac{H}{g} \frac{\partial W}{\partial z_p} \end{aligned}$$

Inserting the above terms into (13.2) and taking into consideration that

$$\cos E = \frac{1}{g_p \gamma_a} \left[\frac{\partial U}{\partial x} \frac{\partial W}{\partial x_p} + \frac{\partial U}{\partial y_a} \frac{\partial W}{\partial y_p} + \frac{\partial U}{\partial z_a} \frac{\partial W}{\partial z_p} \right] \quad (13.3)$$

it is found that (within the accuracy of terms in H^2)

$$U_p = U_a - H \gamma_a \cos E \quad (13.4a)$$

The potentials on the spheroidal and physical surfaces are denoted by U_a and W_p respectively. In figure 13.1 it is assumed that

$$U_a = W_p = C$$

At point P, the spheroid potential, \bar{U} , has the value (eq. 13.4a)

$$U_p = C - H \gamma_a \cos E \quad (13.4b)$$

where the constant C has been introduced. Now substitute $W_p = C$ in equation (13.1) to obtain

$$U_p = C - T_p$$

and therefore, comparing the last two equations, we have

$$T_p = H \gamma_a \cos E \quad (13.5)$$

Equation (13.5) defines Brun's theory which is fundamental in analysis of the gravipotential. The term $\gamma_a \cos E$ can be approximated so as to give

$$\gamma_a \cos E = \gamma_a - \gamma_a \frac{E^2}{2} \quad (13.6)$$

For small E, T_p , in equation (13.5), can be approximated with

$$T_p = H \gamma_a \quad (13.7)$$

Equation (13.5) permits the computation of undulations of the physical equipotential surfaces with respect to the corresponding spheroid of reference. In figure (13.2), the significance of T_p and H are indicated.

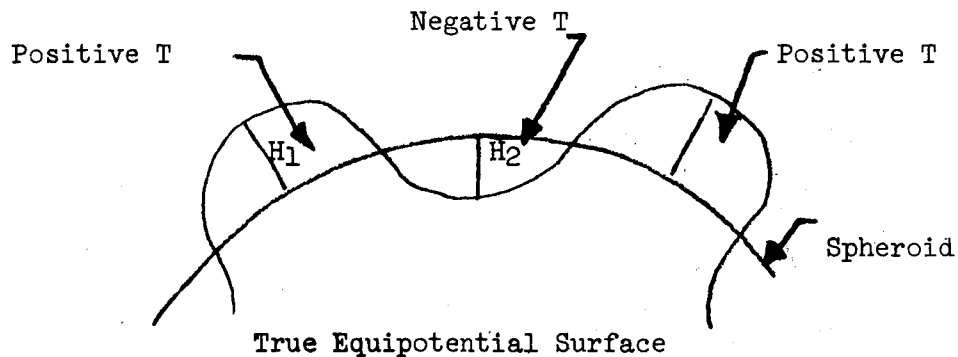


Fig. 13.2

The upwards undulation corresponds to $T > 0$.

In the above analysis, the deduction of Equation (13.6) was based upon a derivation along the normal to the spheroid.

It is of interest to relate the observed gravity, g_p , at a point P to the corresponding theoretical (spheroidal) gravity at this point, g'_p , referred to the relevant spheroid. The difference $g_p - g'_p$ is described as the true gravity anomaly.

Differentiating Equation (13.1) along the geoidal plumb line n_a (figure 13.1) we obtain:

$$\frac{\partial W_p}{\partial n_a} = \frac{\partial U}{\partial n_a} + \frac{\partial T_p}{\partial n_a}$$

Since

$$\frac{\partial W_P}{\partial n_Q} = -g_P \text{ and } \frac{\partial U_P}{\partial n_Q} = \frac{\partial U_P}{\partial n_P} \frac{\partial n_P}{\partial n_Q} = -\gamma'_P \cos E'$$

where, by definition we have $\frac{\partial n_P}{\partial n_Q} = \cos E'$

$$-g_P = -\gamma'_P \cos E' + \frac{\partial T_P}{\partial n_Q}$$

For small angle $E < 1$ min. of arc

$$g_P - \gamma'_P = \gamma'_P \frac{E'^2}{2} - \frac{\partial T_P}{\partial n_Q} \quad (13.7a)$$

or, on first approximation,

$$g_P - \gamma'_P = -\frac{\partial T_P}{\partial n_Q} \quad (13.7b)$$

The error of approximation in equations (13.7a) and (13.7b) at sea level (assuming $\gamma=980$ gal, and $E=1$ min. of arc $= 29 \times 10^{-5}$ radians) is about 0.04 mgal and hence it is, in this case, not significant.

As the next problem, we consider the relation between the gravity at a point, P, on the physical surface, and the corresponding point Q on the spheroid. The potential U_P can be expressed as a power series of the argument H and the potential U_Q as the initial term of the series:

$$U_P = U_Q + \frac{\partial U_Q}{\partial n_Q} H \quad (13.8)$$

and equation (13.1) can be rewritten as

$$W_p = U_a + \frac{\partial U_a}{\partial n_a} H + T_p$$

or

$$W_p = U_a - \gamma_a H + T_p \quad (13.9)$$

Differentiating the last equation along the nadir direction n_a , we obtain for $n_p = \text{spheroidal nadir}$,

$$-g_p = -\gamma_a \cos E - H \frac{\partial \gamma_a}{\partial n_p} \cos E + \frac{\partial T_p}{\partial n_a} \quad (13.10a)$$

For $\cos E \approx 1$,

$$g_p = \gamma_a + H \frac{\partial \gamma_a}{\partial n_a} - \frac{\partial T_p}{\partial n_p} \quad (13.10b)$$

Comparing equations (13.7b) and (13.10b) we obtain,

$$\gamma'_p - \gamma_a = -H \frac{\partial \gamma_a}{\partial n_a} \quad (13.11)$$

From equation (13.5), assuming $\cos E \approx 1$, the elevation difference H can be inserted into (13.10b) yielding a partial differential equation,

$$\left(\frac{\partial T}{\partial n} \right)_p - \frac{T}{\gamma_a} \left(\frac{\partial \gamma_a}{\partial n_a} \right) + \Delta g = 0 \quad (13.12)$$

where $\Delta g = g_p - \gamma_a$

To estimate the value (13.11), we assume as a first approximation

$$\gamma_a = \frac{GM}{r^2}$$

and obtain for $\frac{\partial \gamma_a}{\partial n_a}$ the expression

$$\frac{\partial \gamma_a}{\partial n_a} = \frac{\partial \gamma_a}{\partial r} = -2 \frac{\gamma_a}{r}$$

where r is the geocentric radius vector. Evaluated at sea level, $\gamma_Q = 980 \times 10^3$ mgal, $r = 6.4 \times 10^8$ cm and $H = 50$ m, and the difference of spheroidal gravity values can be computed from equation (13.4) to be

$$\gamma_P - \gamma_Q = 15.6 \text{ mgal}$$

This computation utilized a point on the equator and assumed the uncertainty of the equatorial radius as representing the undulation of the physical sea level.

This result demonstrates that the gravity values γ_P and γ_Q cannot be considered as equal, and shows the need for the evaluation of the physical (geoidal) effect, as given by equation (13.11) for the representation of the anomalous gravity field. The correction given by equation (13.11) is often called the "free air correction." The estimates of the above difference $\gamma_P - \gamma_Q$, shows that any accurate gravity survey requires the free-air correction.

III

THE EARTH'S MAGNETIC FIELDS

14. The Field of a Magnetic Dipole

Electrical charges flowing in a circular ring create a magnetic dipole field. The line integral of the magnetic induction around a closed circuit enclosing a current, I , satisfies the equation

$$I = \oint \frac{\bar{B}}{\mu} \cdot d\bar{s} = \iint \bar{J} \cdot d\bar{A} \quad (14.1)$$

where \bar{B} is the magnetic induction; μ is the permeability; $d\bar{s}$ is a length increment vector; \bar{J} is the current density per unit area; and \bar{A} is the axial vector representing the area enclosed by the line integration. The field is symmetrical about the z axis in figure 14.1; and, therefore, it is a two dimensional problem.

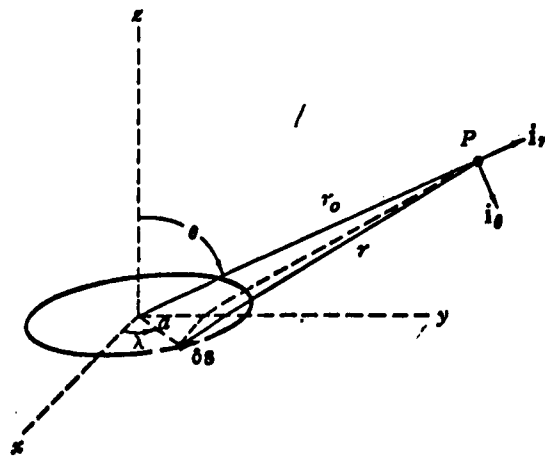


Fig. 14.1-Magnetic Dipole

An electrical current flowing in the line element $d\bar{s}$ creates a magnetic induction field at P. Due to the geometry of the problem, the magnetic field is directed along the vector \bar{l}_θ , perpendicular to the vector \bar{r}_0 ; and the gradient of the potential is in the direction θ . θ is the colatitude. The total derivative of the scalar magnetic potential is given by the relation

$$dV = \frac{\partial V}{\partial \theta} d\theta = \bar{l}_\theta \cdot \left[\int r d\bar{B} \right] d\theta \quad (14.2)$$

where $d\bar{B}$ is the differential increment of field produced by the x-component of the current in $d\bar{s}$. The expression for the differential, $d\bar{B}$, in the integral can be expanded to give for the integral

$$\int r d\bar{B} = \left(\frac{\mu}{4\pi} \right) \bar{I} a \int_0^{2\pi} \frac{\bar{l} \times \bar{l}_r}{r} \sin \lambda d\lambda \quad (14.3)$$

Since $r^2 = r_0^2 + a^2 - 2r_0 a \sin \theta$ for $r \gg a$, only the first two terms of the series expansion need be included in the expression for $\frac{1}{r}$:

$$\frac{1}{r} = \frac{1}{r_0} \left(1 + \frac{a \sin \theta \sin \lambda}{r_0} \right) \quad (14.4)$$

After two integrations, the magnetic scalar potential becomes

$$V = - \frac{M}{r_0^2} \cos \theta \quad (14.5)$$

where the magnetic moment, M , stands for

$$M = \left(\frac{\mu}{4\pi} \right) \bar{I} \pi a^2 \quad (14.6)$$

Taking the gradient of the potential, the magnetic induction can be expressed as

$$\vec{B} = \vec{i}_r \frac{2M \cos \theta}{r^3} + \vec{i}_\theta \frac{M \sin \theta}{r^3} \quad (14.7)$$

This equation approximates the well known geomagnetic field. The dipole field of the center can be calculated and it is illustrated on Figure 14.2.

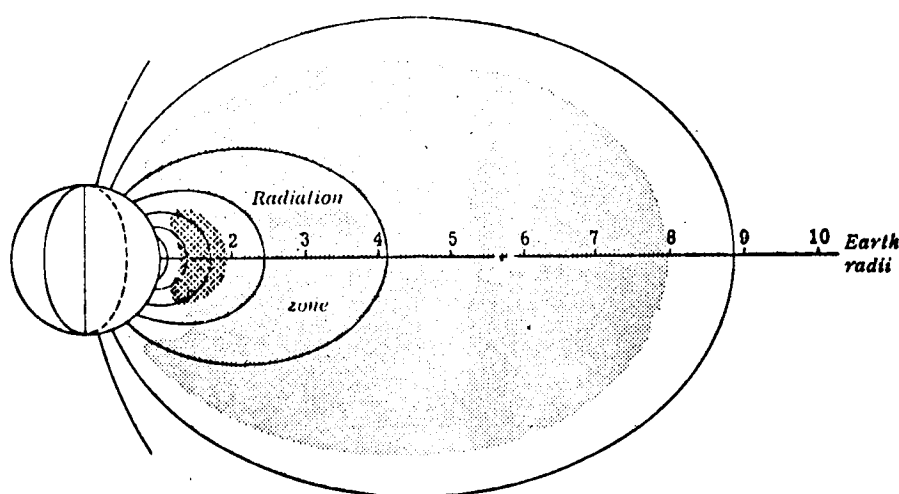


Fig. 14.2-Earth's Magnetic Field

Source: Fleagle and Businger, An Introduction to Atmospheric Physics, Academic Press, New York and London, 1963

15. Earth's Magnetic Field

The magnetic field of the Earth can be approximated by a dipole at the center of the earth. But a closer examination of the magnetic field discloses significant departures from a dipole field; these departures are called geomagnetic anomalies. The lines of force leave the northern hemisphere and re-enter the Earth in the southern hemisphere. A freely balanced needle will align itself along the line of force passing through it. The needle will be orientated vertically at the magnetic pole, and horizontally at the magnetic equator. As we all know, the magnetic poles and equator of the Earth

do not coincide with the geographical ones. It should be noticed that the north magnetic pole is of "south" polarity. The magnetic force, \vec{B} , describes the magnetic field at any point. It can be given in terms of three different sets of parameters:

- (a) D = declination H = horizontal intensity z = vertical intensity
- (b) D = declination H = horizontal intensity I = inclination = dip
- (c) X = N-S component Y = E-W component Z = vertical component

The components and their signs are indicated in figure 15.1

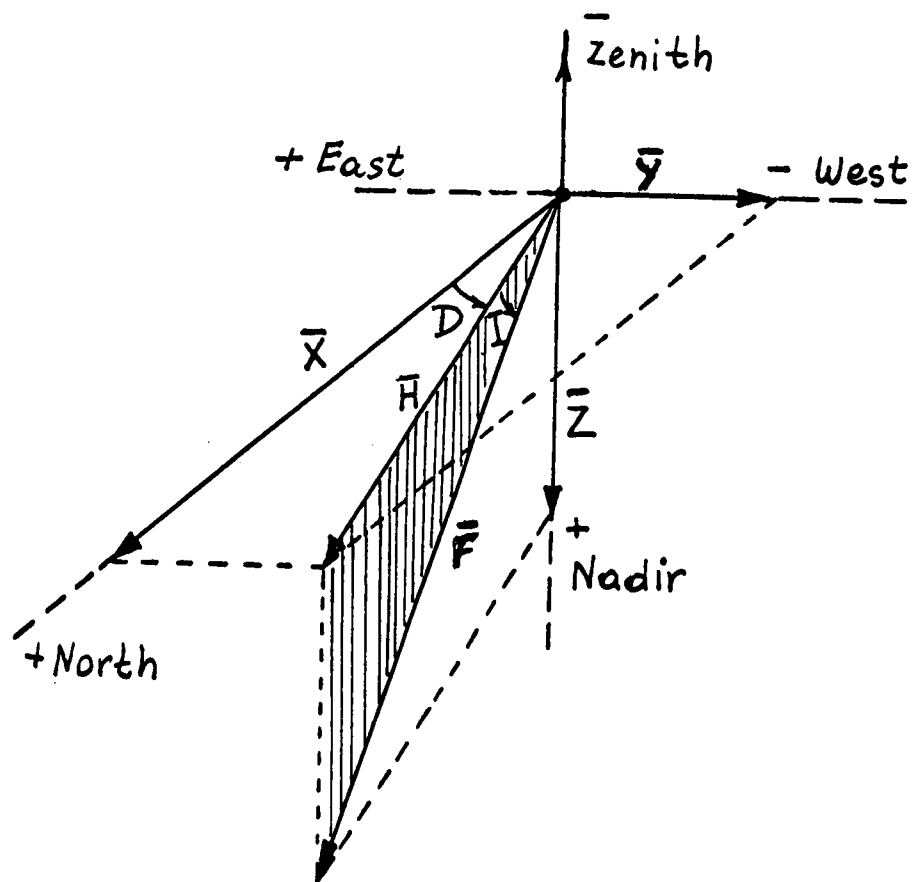


Fig. 15.1-Components of Earth's Magnetic Force

The scalar potential of a magnetic dipole at the colatitude angle θ (equal to $90^\circ - \phi$) is defined by equation (14.5) as

$$V = - \frac{M}{r^2} \cos \theta \quad (15.1)$$

where the magnetic moment for the earth is

$$M = 8.06 \times 10^{25} \text{ gauss cm}^3$$

The vector field of a dipole is obtained from the scalar potential by taking the gradient:

$$\vec{B} = - \nabla V \quad (15.2)$$

It is customary to describe the geomagnetic field on the earth's surface by expanding the magnetic scalar potential in a spherical harmonic series. The origin of the coordinates is placed at the earth's geometric center and the polar axis of the spherical polar coordinate system is assumed to be parallel to the earth's geographic axis. The expansion takes a form similar to that for the gravitational potential:

$$V = \sum_{n=0}^{\infty} \left(\frac{a}{r}\right)^{n+1} \sum_{m=0}^n P_{nm}(\cos \theta) (g_{nm} \cos m\lambda + h_{nm} \sin m\lambda) \quad (15.3)$$

where a is the earth's radius, r is the distance from the earth's center; $\theta = 90^\circ - \phi$ is the geographic colatitude; λ is the ^{east} longitude of the point; $P_{nm}(\cos \theta)$ is the normalized Legendre polynomial of the n^{th} degree and m^{th} order; g_{nm} and h_{nm} are gaussian coefficients determined on the basis of physical measurements. This representation can be used only where currents are not present. Successful applications of the formula seem to prove

that the external currents are small and cannot produce more than 1/1000 of the field at the earth's surface.

Since the geomagnetic field is subject to secular changes, the gaussian coefficients must be reevaluated from time to time. The comparison of world magnetic charts for the surface field (figs. 15.2 and 15.3) for the decade 1945 to 1955 show that significant changes in the total intensity have occurred (H.E. Vestine, 1964), total intensities are given in gauss.

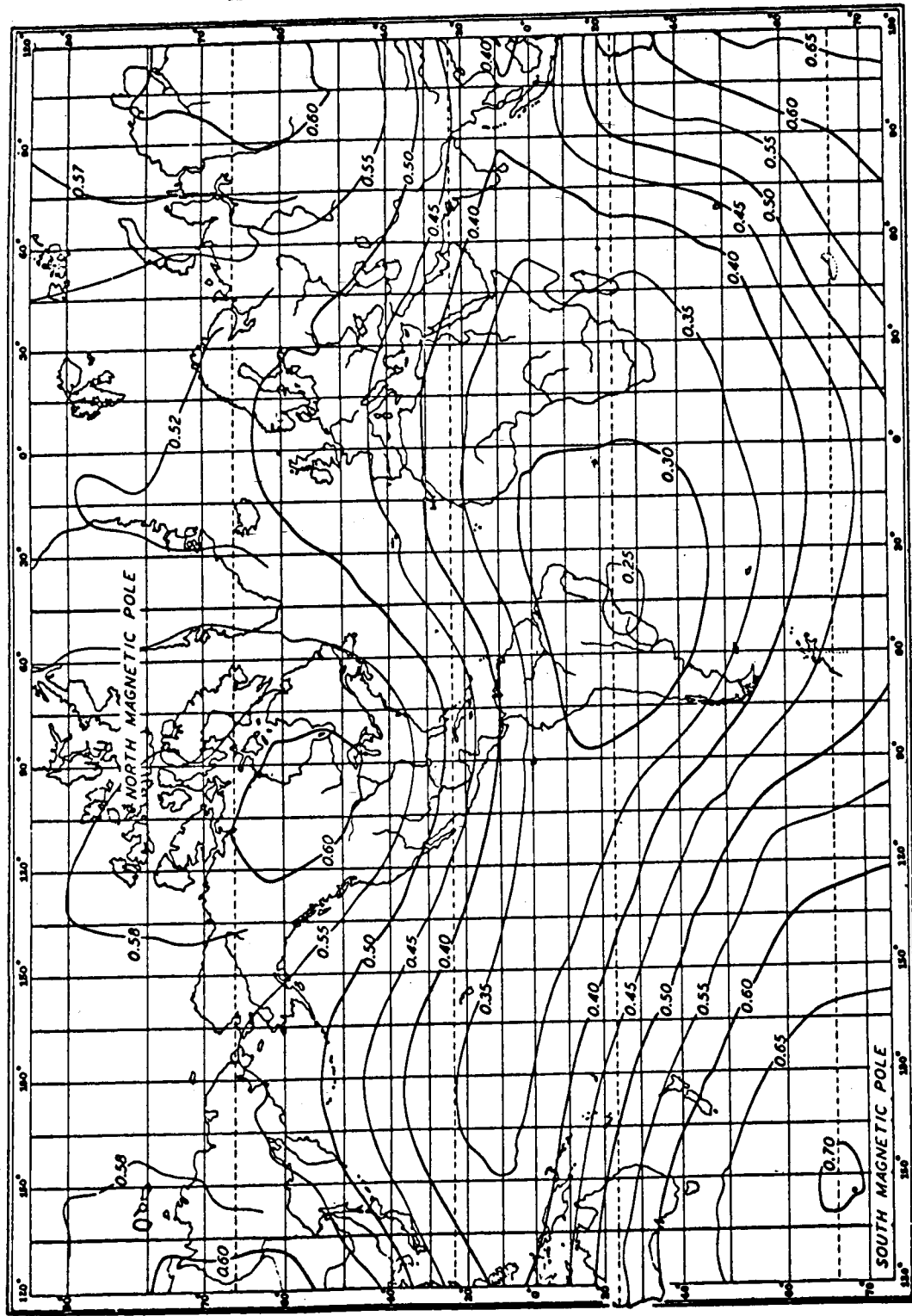


Fig. 15.2-World magnetic chart of the surface field: total intensity, 1945
(H. E. Vestine, The RAND Corp. Publ. F-2978, 1964)

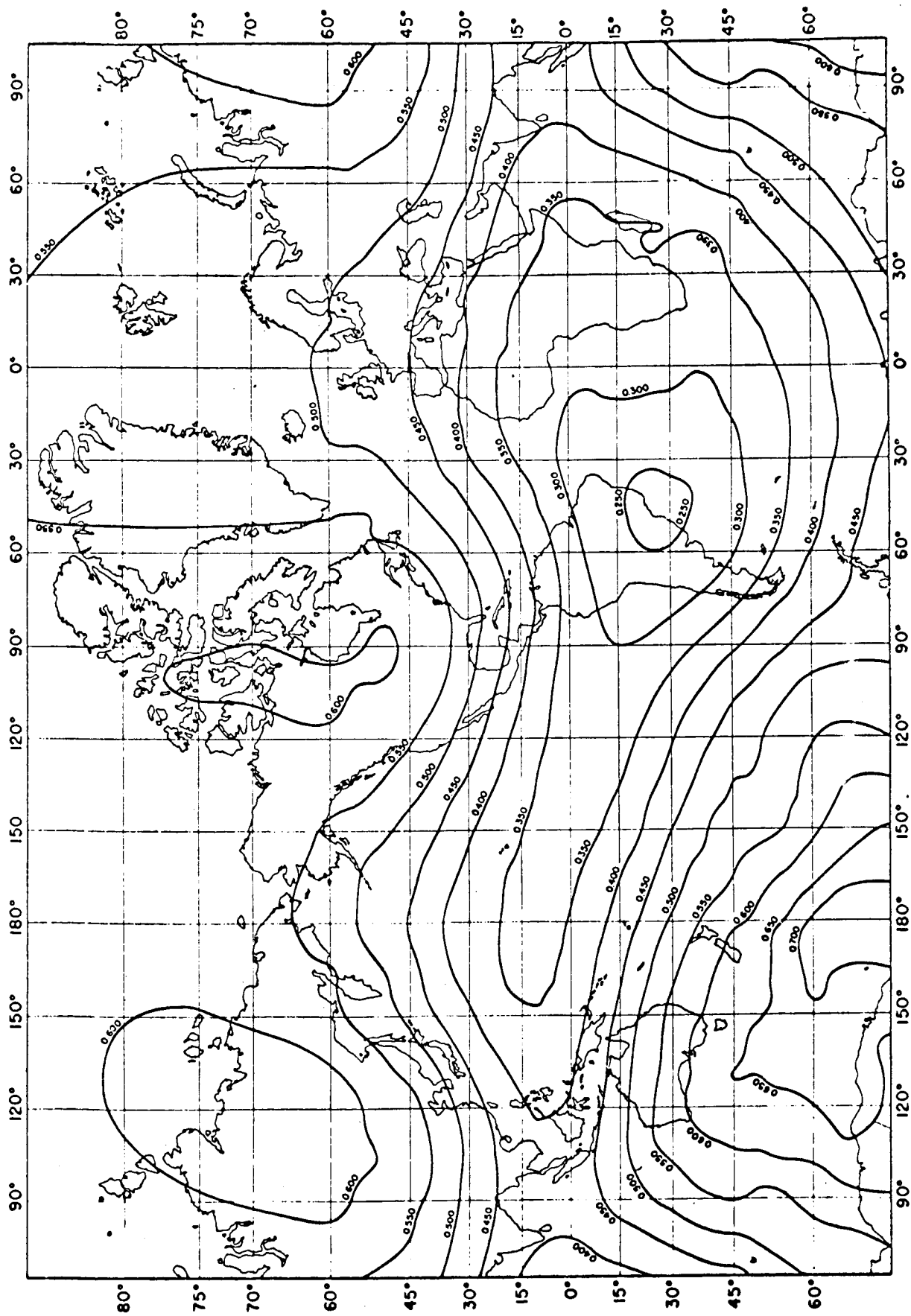


Fig. 15.3-World Magnetic Chart of the Surface Field;
total intensity, 1955 (H.E. Vestine, The RAND Corp.
Publ, F-2978, 1964)

The geomagnetic field is also affected by solar activity, by variations ^{variations (S_q)} in the solar quiet, which depend on the time of solar day; by various disturbances which occur from day to day (S_D), and by lunar daily (L) variations occurring together with solar quiet variations (S_q). All these perturbations are depending on the time of year 11-year solar cycle and the geomagnetic latitude.

Variations of the geomagnetic field are correlated to large scale current systems in the ionosphere between 100 and 150 Km. It is believed that the currents are generated by daily heating and cooling processes caused by the sun. These processes are responsible for transporting ions and electrons across field lines.

16. Shifts in the Magnetic Poles

Analysis of old lavas and other geological specimens indicate that reversals of the earth's magnetic field have occurred. The study made by Columbia's Lamont Geological Observatory shows that the reversal of polarity takes place at intervals of from half a million years to a million years; the changes take about 10,000 years to be completed.

During the reversal process, the intensity of the magnetic field reaches zero and then build up again with the opposite polarity. Since the magnetic field shields the earth from cosmic rays, and thus protects life on earth, some species of life must be killed off and others must undergo mutations during periods of weak magnetic field. Thus, the reversals of polarity must have had some influence on the evolution of life on this planet.

Oceanic samples indicate that the last reversal occurred approximately 0.7 million years ago; temporary reversals took place 0.9 and 1.9 millions of years ago; two other semi-permanent changes of polarity took place 2.4

and 3.5 millions of years ago.

At the present time the intensity of the magnetic field is decreasing; and, if the present rate of decrease in intensity continues, the intensity will reach zero in about another 2,000 years. In addition, a northeast shift in the north magnetic pole has been observed in the last century. The position of the pole was first established in 1831. At that time, it was located off the coast of the Boothia Peninsula north of King William Island. By 1904, it had moved northeast about 25 miles. During the next 44 years it moved about 250 miles to a point in the Barrow Strait, north of Prince of Wales Island. From 1948 to 1962 the pole had moved about 80 miles northeast to a point near Peddie Bay at the southern end of Bathurst Island. In the last two years the pole has moved 20 miles north and four miles east.

Project Mohole, designed to drill into the ocean mantle, will attempt to obtain additional records of magnetic reversals in the Pacific, Atlantic, and Indian oceans. The lavas and sediments will be analyzed for when the successive magnetic reversals took place.

17. Deformation of Earth's Magnetic Field in Space

The geomagnetic field would extend to infinity in the complete vacuum of interplanetary space. For a long time geophysicists have observed that after the appearance of solar flares the magnetic storms occur. Based on this observation S. Chapman and V. C. A. Ferraro (1930) proposed that plasma clouds produced by solar flares deform the geomagnetic fields, and cause magnetic storms.

Piddington (J.G.R. 65, 93, 1960) suggested that the magnetic field be investigated by satellites to determine whether magnetic lines are extended on the night side and compressed on the day side of the earth. See figures 17.1 and 17.2 (from N. F. Ness, Science, vol. 151, no. 3714, p. 1042 and p. 1046).

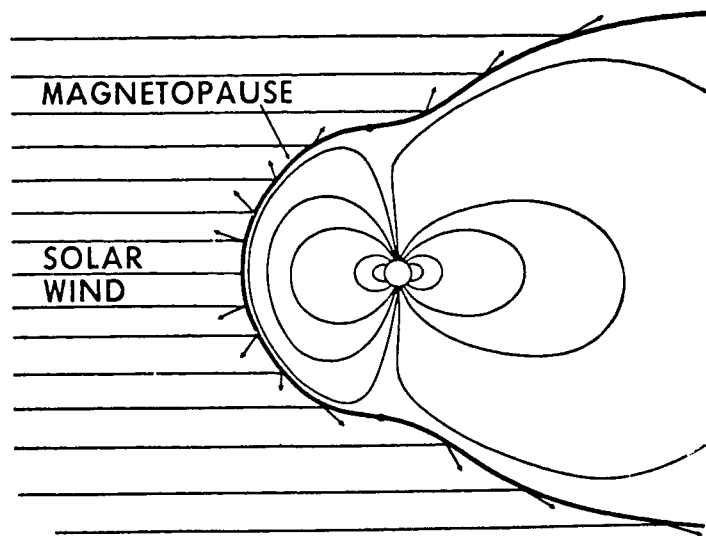


Fig. 17.1-Simplified representation of the interaction of the rarified solar wind plasma with the geomagnetic field.

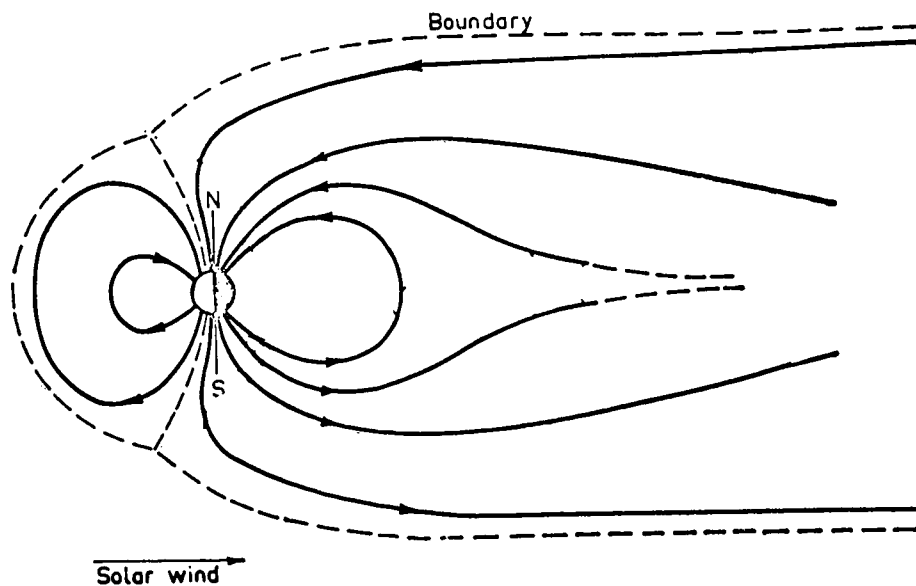


Fig. 17.2-Schematic diagram of the geomagnetic field as distorted by the solar wind.

Direct measurements in space indicate that solar wind velocities are in the range of 3 to 7×10^2 kilometers per second and the densities of protons are in the range of 3 to 70 protons per cubic centimeter.

The solar wind compresses the magnetic field into a rounded shell on the day side to a thickness of about $40,000$ miles. There is disagreement as to how long the magnetic tail is on the night side of the earth. Mariner IV reported the absence of electron fluxes at a distance of 3000 earth radii. Recent magnetic measurements by IMP-I and Luna 10 have shown that the magnetic field extends to the distance of the Moon's orbit on the night side of the earth (see figure 17.3). Luna 10 measured the presence of electrons in the vicinity of the Moon with intensities far greater than anticipated. It appears that the electrons are confined there by the earth's magnetic tail, in much the same way as the electrons and protons are trapped by the earth's magnetic field to form the Van Allen radiation belts.

The charged particles in the vicinity of the moon induce weak electric currents within the moon. These currents in turn generate a weak magnetic field in the moon's vicinity.

There is some similarity between the magnetic tail and the tail of a comet. The analogy is merely descriptive and it is up to future experimental and theoretical research to prove or disprove the analogy.

Compared to the earth's magnetic moment of 8×10^{25} c.g.s. units, the magnetic fields of Venus and Mars, as indicated by planetary probes, are 3.4 percent and 0.03 percent respectively. Results of satellite experiments are essential in determining the origin and history of the geomagnetic field. They must be considered along with paleomagnetic evidence.

18. Charged Particles in the Earth's Electromagnetic Field

In general, electrons and positive ions move in a plane which is inclined to the direction of the magnetic field. Assuming that the y -axis coincides with

element to the covered rows and subtract it from the uncovered columns (or add it to the covered columns and subtract it from the uncovered rows) (or add it to the twice-covered elements and subtract it from the uncovered elements). Do not change any stars, primes, or coverings. Go to step 3.

The original problem is shown in figure 11.

2	6	5	9
3	4	8	8
5	1	2	3
4	3	2	7

Figure 11

Subtract from each row its smallest element (figure 12)

0	4	3	7
0	1	5	5
4	0	1	2
2	1	0	7

Figure 12

Subtract from each column its smallest element (figure 13)

0	4	3	5
0	1	5	3
4	0	1	0
2	1	0	5

Figure 13

the lines of magnetic field, the velocity component in the x-y plane can be written as

$$v_e = v_0 \sin \alpha$$

where α is the angle between the particle's velocity vector and the direction of the magnetic field. The equation for the magnetic force, F , can be written in the form

$$\frac{\vec{F}}{e} = \vec{v}_e \times \vec{B} \quad (18.1)$$

where e is the electric charge; \vec{v}_e is the velocity of the charge; and \vec{B} is the magnetic induction. According to this equation, positive ions are deflected counterclockwise, while electrons are deflected clockwise, when looking toward the north, as indicated in figure 18.1.

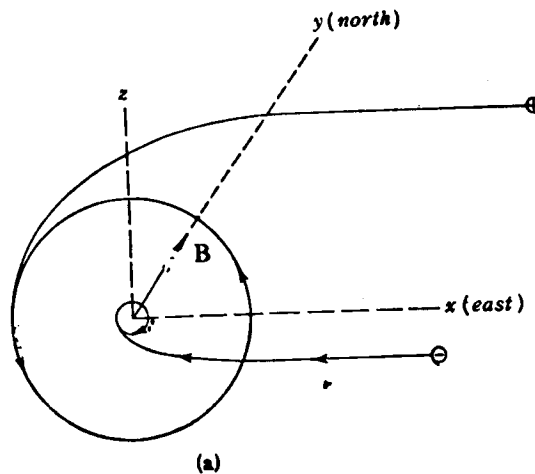


Fig. 18.1-Deflection of an electron (-) and proton (+)

IV

RADIATIVE PROCESSES IN THE ATMOSPHERE

There are various ways of describing the structure of the atmosphere: according to chemical reactions, ionization, composition, temperature, and molecular escape. See figure IV. The standard nomenclature accepted by the International Union of Geodesy and Geophysics is defined in Table IV. This section is concerned with the composition of the atmosphere. The description begins at the Earth's surface and progresses upward to the junction of the atmosphere with the interplanetary gas. The terminus of complete mixing of atmospheric constituents at approximately 100 kilometers serves as a convenient division of the atmosphere into two levels. Below 100 kilometers the prevailing process is the mixing of gases in the form of neutral molecules. The mixture consists mainly of molecular oxygen and nitrogen in the considerable, variable amounts of water vapor. The concentration of water vapor is greatest near the surface of the Earth due to condensation.

Above 100 kilometers the diffusive process dominates. As a result, various gases occur in their atomic form and stratify at various levels according to their atomic weight. At this height gases dissociate into negative and positive ions.

ALTITUDE		IUGG RECOMMENDED NOMENCLATURE					GERSON	GOODY
		CHAPMAN				SPITZER		
		CHEMICAL REACTIONS	IONIZATION	COMPOSITION	TEMPERATURE			
MILES	KILOMETERS					MOLECULAR ESCAPE		
600	1000					EXO-SPHERE	EXO-SPHERE	EXO-SPHERE
500								
400				METRO-SPHERE	THERMO-SPHERE		MESO-SPHERE	
300	500							
200	300							
100	200		IONO-SPHERE				IONO-SPHERE	IONO-SPHERE
50	100							
30	50	CHEMO-SPHERE		HOMOPAUSE HOMO-SPHERE	MESOPAUSE MESO-SPHERE		CHEMO-SPHERE	STRATO-SPHERE
20	30							
10	20				STRATOPAUSE STRATOSPHERE		STRATOSPHERE	
8	10				TROPOPAUSE			
1	1				TROPO-SPHERE		TROPO-SPHERE	TROPO-SPHERE

Fig. IV.--Systems of Nomenclature of Atmospheric Shells
 Science: Handbook of Geophysics USAF Rev. Ed.
 Mac Millan , 1961

Table IV IUGG Description of Atmospheric Shells

Based on Temperature

Troposphere The region nearest the surface, having a more or less uniform decrease of temperature with altitude. The nominal rate of temperature decrease is 6.5° K/km, but inversions are common. The troposphere, the domain of weather, is in convective equilibrium with the sun-warmed surface of the earth. The tropopause, which occurs at altitudes between 6 and 18 kilometers (higher and colder over the equator), is the domain of high winds and highest cirrus clouds.

Stratosphere The region next above the troposphere and having a nominally constant temperature. The stratosphere is thicker over the poles, thinner or even nonexistent over the equator. Maximum of atmospheric ozone found near stratopause. Rare nacreous clouds also found near stratopause. Stratopause is at about 25 kilometers in middle latitudes. Stratospheric temperatures are in the order of arctic winter temperatures.

Mesosphere	The region of the first temperature maximum. The mesosphere lies above the stratosphere and below the major temperature minimum, which is found near 80 kilometers altitude and constitutes the mesopause. A relatively warm region between two cold regions; the region of disappearance of most meteors. The mesopause is found at altitudes of from 70 to 85 kilometers. Mesosphere is in radiative equilibrium between ultraviolet ozone heating by the upper fringe of ozone region and the infrared ozone and carbon dioxide cooling by radiation to space.
Thermosphere	The region of rising temperature above the major temperature minimum around 80 kilometers altitude. No upper altitude limit. The domain of the aurorae. Temperature rise at base of thermosphere attributed to too infrequent collisions among molecules to maintain thermodynamic equilibrium. The potentially enormous infrared radiative cooling by carbon dioxide is not actually realized owing to inadequate collisions.

Based on Composition

Homosphere	The region of substantially uniform composition, in the sense of constant mean molecular weight, from the surface upwards. The homopause is found at altitudes between 80 and 100 kilometers. The composition changes here primarily because of dissociation of oxygen. Mean molecular weight decreases accordingly. The ozonosphere, having its peak concentration near stratopause altitude does not change the mean molecular weight of the atmosphere significantly.
Heterosphere	The region of significantly varying composition above the homosphere and extending indefinitely outwards. The "molecular weight" of air diminishes from 29 at about 90 kilometers to 16 at about 500 kilometers. Well above the level of oxygen dissociation, nitrogen begins to dissociate and diffusive separation (lighter atoms and molecules rising to the top) sets in.

Based on Ionization

Ionosphere	The region of sufficiently large electron density to affect radio communication. However, only about one molecule in 1000 in the F ₂ region to one in 100,000,000 in the D region is ionized. The bottom of the ionosphere, the D region, is found at about 80 kilometers during the day. At night the D region disappears and the bottom of the ionosphere rises to 100 kilometers. The top of the ionosphere is not well defined but has often been taken as about 400 kilometers. The recent extension upward to 1000 km based on satellite and rocket data is shown.
------------	---

Based on chemical Reactions reactions

Chemosphere The region where chemical activity (primarily photochemical) is predominant. The chemosphere is found within the altitude limits of about 20 to 110 kilometers.

Based on Molecular Escape escape

Exosphere The region wherein molecular escape from the earth's atmosphere is significant. The base of the exosphere, the critical level, is thought to be at an altitude above 300 kilometers, possibly as high as 1000 kilometers. Satellite data indicating higher densities at these altitudes favor higher exosphere levels. Lighter atoms and molecules can escape at lower altitudes than heavier ones. The earth's magnetic field effectively prevents the escape of charged particles, however.

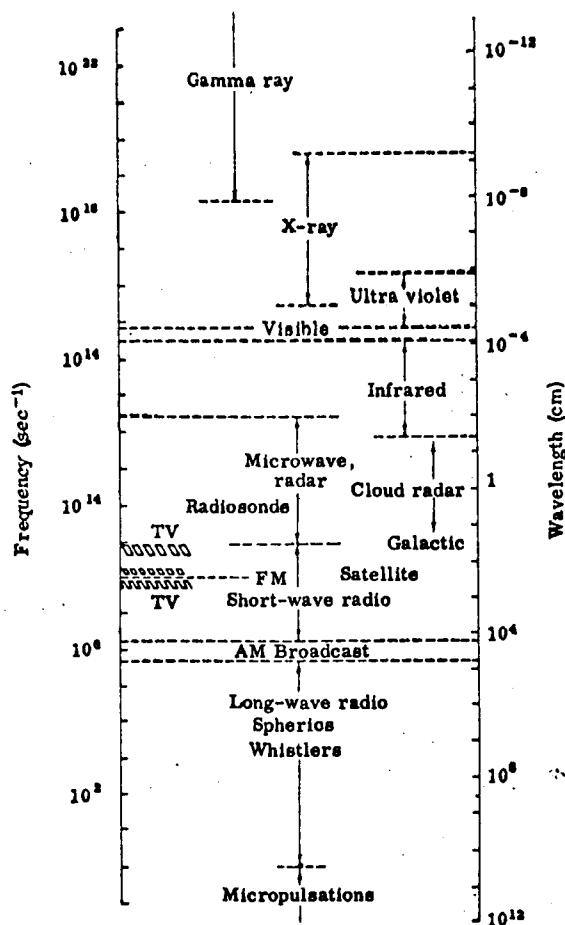
Source: Handbook of Geophysics USAF Rev. Ed. Mac Millan, 1961

The planetary atmosphere is formed by the evaporation of gases from Earth's crust, and held by the gravitational force of the planetary mass. This explains the fact that the density of the atmosphere and its vertical distribution depend on the mass of a planet. Additional factors involved in the formation of planetary atmosphere are: the chemical reaction between the component gases and the crust, the escape gases, and the heating and photochemical processes caused by solar radiation entering the atmosphere.

The sun is the principal source of energy causing atmospheric processes. All other sources of energy are negligible. Solar energy is in part reflected, in part absorbed or transmitted by the atmosphere. A portion of the radiative energy absorbed by the atmosphere is reemitted again and either escapes into space or reaches the surface of the Earth.

Solar radiation either directly or indirectly reaches the earth. Some energy consists of both direct and scattered solar radiation; some energy is reemitted by the atmosphere.

The nomenclature of the electromagnetic spectrum is given below (Fleagle and Businger, An Introduction to Atmospheric Physics, Associated Press, New York, 1963.)



The spectral characteristics of radiant energy depend upon its origin. Solar radiation is predominantly short-wave radiation, shorter than 3-4 microns. The emission and the atmosphere is concentrated in the long-wave portion of the spectrum, longer than 2 microns.

19. Definitions of Radiation Field

The fundamental concept of the radiation field is the radiant flux the density of radiant flux is defined as the amount of radiant energy (E) received from all directions per unit time (dt) and per unit area (dA).

$$F = \frac{d^2 E}{dA dt} \quad (19.1)$$

cal cm⁻² min⁻¹

in units of cal cm⁻² min⁻¹

The intensity is defined as radiant energy per unit time coming from a specific direction and passing through a unit area perpendicular to the direction of the incoming radiation. See Figure 19.1.

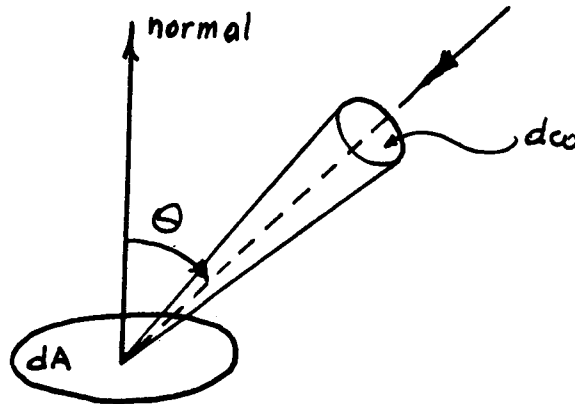


Fig. 19.1-Incoming radiation

According to this definition, the intensity can be expressed as

$$I = \frac{dF}{d\omega \cos \theta} \quad (19.2)$$

and the total flux density can be expressed as

$$F = \int_0^{2\pi} I \cos \theta d\omega$$

In certain wave length intervals, λ and $\lambda + d\lambda$, the flux density can be written as

$$dF_\lambda = I_\lambda \cos \theta d\omega dt d\lambda dA = \frac{dF}{d\lambda} \quad (19.3)$$

where I_λ = monochromatic intensity

$$I_\lambda = \frac{dI}{d\lambda}$$

and also

$$I = \int_0^\infty I_\lambda d\lambda \quad (19.4)$$

The intensity is expressed in cal/cm²min sterad. Flux intensity in all wave lengths can also be given by the equation:

$$F = \int_0^\infty F_\lambda d\lambda$$

When the intensity of the radiation does not depend upon the direction of the incoming energy, the radiation is called isotropic radiation. According to Figure 19.2 the solid angle $d\omega$ is defined as:

$$d\omega = \frac{r \sin \theta d\phi \cdot r d\theta}{r^2} = \sin \theta d\theta d\phi$$

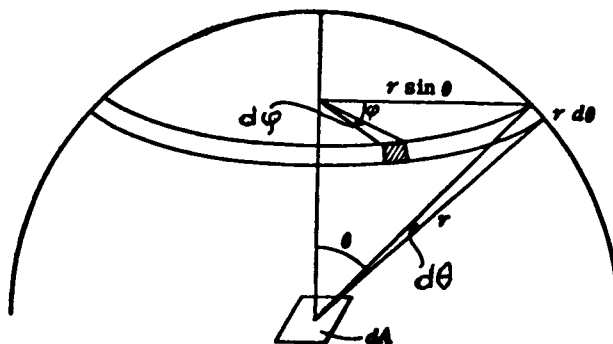


Fig. 19.2-Geometry of solid angle

By introducing the expression for dw , we obtain

$$dF_{\lambda} = I_{\lambda} \cos \theta \sin \theta d\theta d\phi dt dA d\lambda \quad (19.5)$$

and between taking $dA=1$ and $dt=1$, the flux density

$$F_{\lambda} = \int_0^{2\pi} d\phi \int_0^{2\pi} I_{\lambda} \cos \theta \sin \theta d\theta \quad (19.6)$$

In the case of isotropic radiation, $I_{\lambda} = \text{constant}$.

$$F_{\lambda} = I_{\lambda} \int_0^{2\pi} d\phi \int_0^{2\pi} \cos \theta \sin \theta d\theta = \pi I_{\lambda} \quad (19.7a)$$

and the total flux

$$F = \int_0^{\infty} F_{\lambda} d\lambda = \pi I \quad (19.7b)$$

The equation (19.7b) can be described in such a way that the flux of isotropic incoming radiation from a hemisphere received by a surface is equal to the product of π times the intensity. The surface can be arbitrarily oriented.

20. Absorption Reflection and Emission of Radiation

Radiant energy can be absorbed in a medium, reflected from a medium or transmitted through a medium. Let us denote the ratios of the absorbed, reflected and transmitted energy to the incident radiation by a_{λ} , r_{λ} and τ_{λ} ; the following relationship exists between these ratios.

$$a_{\lambda} + r_{\lambda} + \tau_{\lambda} = 1 \quad (20.1)$$

A mass element dm emits in all directions the same amount of energy dE_λ

$$dE_\lambda = e_\lambda dm d\omega d\lambda \quad (20.2)$$

where e_λ = mass emission coefficient. The total emission coefficient in all wave lengths is:

$$e = \int_0^\infty e_\lambda d\lambda$$

The narrow energy beam of intensity I_λ passes through a medium of density ρ along the path dx . Absorption is proportional to the path length, density of the medium and intensity of the incident radiation

$$dI_\lambda = -k_\lambda \rho I_\lambda dx \quad (20.3)$$

where k_λ = the absorption coefficient given in cm^2g^{-1} and density ρ is given in g cm^{-3} . By integrating equation (20.3) we get

$$I_\lambda = I_{\lambda 0} \exp \left(-k_\lambda \int_0^x \rho dx \right) \quad (20.4)$$

This equation is known as Beer's law. The quantity

$$v = \int_0^x \rho dx$$

is called the optical thickness.

As mentioned before, a portion of the incident radiation is reflected back, this amount, depending upon the reflective property of the body. Diffuse reflection is called albedo, and it is defined as the ratio of the reflected flux to the incident flux. It should be pointed out that reflection is also a function of wave lengths.

21. Laws of Emission

According to Kirchhoff's law, the ratio of the emission coefficient e_λ to its absorption coefficient k_λ is a function of temperature and wave length.

$$e_\lambda : k_\lambda = I(\lambda, T) \quad (21.1)$$

If we consider a black body, $k_\lambda = 1$, the function $I(\lambda, T) = I_{\lambda T}$ the intensity of black body radiation for the same wave length and temperature. This law is valid only for thermodynamic equilibrium where the body emits the same amount of energy as it absorbs. The atmosphere however is not in a state of thermal equilibrium, therefore, in order to apply Kirchhoff's law an assumption as to the local thermal equilibrium must be made.

Planck wrote the equation for the intensity of black body radiation in the form,

$$I_{\lambda T} = \frac{c_1}{\lambda^5} [\exp(c_2/\lambda T) - 1]^{-1} \quad (21.2)$$

($c_1 = 2\pi^5 c^2 h$, where c = velocity of light, $h = 6.625 \times 10^{-27}$ erg sec⁻¹ = Planck's constant; $c_2 = c h/k$, where $k = 1.38 \times 10^{-16}$ erg deg⁻¹ = Boltzmann's constant) Experimentally it was found that $c_1 = 3.74 \times 10^{-5}$ erg/cm² sec and $c_2 = 1.439$ cm deg, while λ in cm and T in K^o.

Figure 21.1 illustrates the computed intensities of black body radiations for the temperatures 5000^oK, 6000^oK, and 7000^oK. (per unit wavelengths).

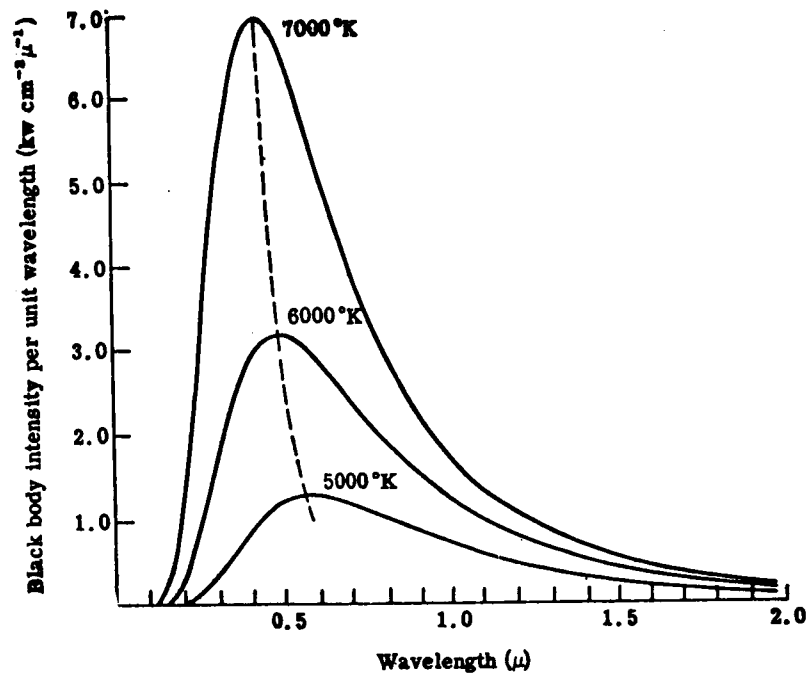


Fig. 21.1-Black body radiation for temperatures 5000°K, 6000°K, and 7000°K

Inspection of equation 21.2 demonstrates that the intensity increases very rapidly with rising temperature and that its peak emission shifts towards shorter wave lengths. The equation can also be given in the following quantized

form.

$$I_{\lambda T} = 2hc^2 \lambda^{-5} [\exp(hc/k\lambda T) - 1] \quad (21.3)$$

where $h = 6.625 \times 10^{-27}$ erg sec = Planck constant, c = velocity of light, and $k = 1.38 \times 10^{-16}$ erg deg $^{-1}$ = Boltzmann constant. According to equation 19.7b the flux of radiation can be represented in terms of the intensity. Thus the total radiation intensity of a black body for all wavelengths will be found as,

$$F_T = \int_0^{\infty} F_{\lambda T} d\lambda = \pi \int_0^{\infty} I_{\lambda T} d\lambda = \frac{2\pi^5 k^4}{15c^2 h^3} T^4 = bT^4 \quad (21.4)$$

This relation is known as the Stefan-Boltzmann equation. Using the Stefan-Boltzmann constant, ^{we find} $b = 5.669 \times 10^{-12} \text{ watt cm}^{-2}\text{deg}^{-4} = 0.814 \times 10^{-14} \text{ cal cm}^{-2} \text{ min}^{-1}\text{deg}^{-4}$.

In order to find the wave length of maximum intensity, we differentiate Planck's law with respect to the wave length and we equate to zero. From this relationship, Wien's displacement law is obtained.

$$\lambda_m = \alpha T^{-1} \quad (21.5)$$

where $\alpha = 0.288 \text{ cm deg}$. This equation also makes it possible to determine the color temperature of the body corresponding to maximum emission in certain wave lengths.

22. Solar Radiation in the Upper Atmosphere

The sun can be approximated by a black body radiating energy at a temperature of 6000°K . Figure 21.1 shows the spectral energy distribution of electromagnetic radiation computed for temperatures of 6000°K and 5700°K (solid line) and observed (dashed line). Source: F. S. Johnson, J. Meteorol, vol. 11, 431, 1951.

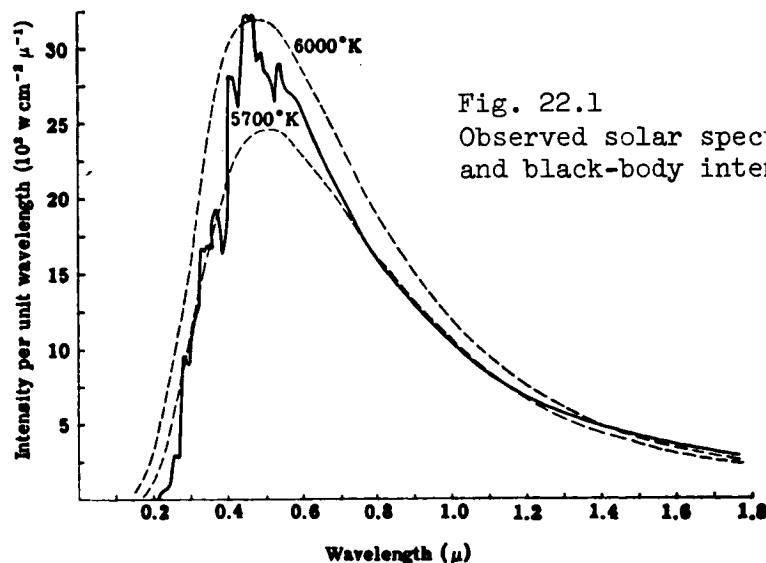


Fig. 22.1
Observed solar spectrum
and black-body intensities .

Short wave solar radiation reaches the earth; the earth in turn re-emits longer wave radiation (infrared), which is absorbed by the atmosphere, and in particular by H_2O and O_2 ; O_3 , N_2O , and CH_4 absorb infrared radiation to a lesser extent. This absorption heats the atmosphere, which re-radiates the energy, a part of it downward; this energy provides additional thermal energy for the earth's surface. This process of trapping solar energy is called the "greenhouse effect," a process in which infrared wave radiation prevails.

Figure 21.2 gives the absorption spectra for water vapor (H_2O), carbon dioxide (CO_2), diatomic oxygen (O_2), and ozone (O_3), nitrous oxide (N_2O), methane (CH_4) and the atmosphere. Source: J. N. Howard, Proc. IRE, vol. 47, 1451 (1959); R. M. Goody and G. D. Robinson, J. Roy. Met. Soc. vol. 77, 153, (1951).

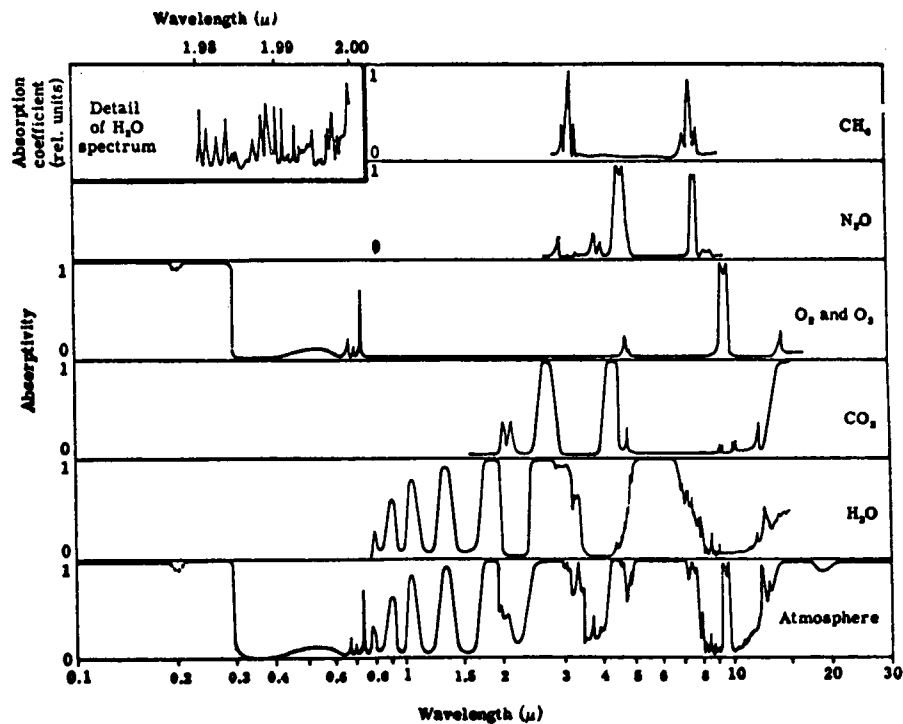
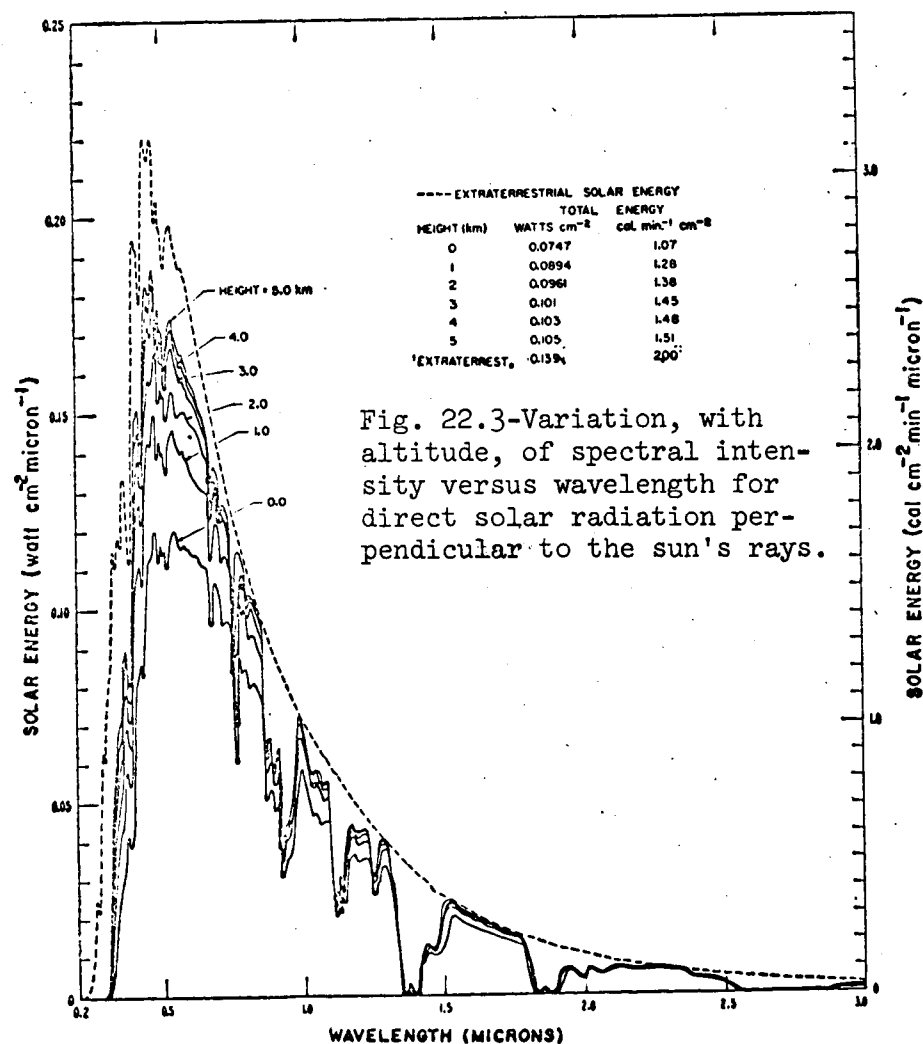


Fig. 22.2-Absorption spectra for H_2O , CO_2 , O_2 , N_2O , CH_4 , and the absorption spectrum of the atmosphere

It should be noted that the radiation in the range between 8 and 12 microns can escape into space-- with the exception of the ozone absorption band at 9.6 microns. This wave interval is called the atmospheric infrared "window."

The density of the atmosphere diminishes with increase in altitude and accordingly there is a decrease in the total number of air molecules, water vapor, carbon dioxide and aerosol. This decrease in molecules changes the transmissivity of the atmosphere, and as a result there is a variation in the distribution of the solar spectrum at various altitudes. Figure 22.3 illustrates this variation of spectral intensity where the air mass is 1.5 and the concentration of precipitable water, 5.0 millimetres; the concentration of aerosol, 200 particles per cubic cm; of ozone, 0.35 cm. Source D. M. Gates, Science, vol. 151, p. 528, 1966.



The constituents of the upper atmosphere have no absorption spectra in the infrared; they are affected by short wave radiation, both near and extreme ultraviolet (EUV) and X rays. The ultra violet, EUV and X rays comprise only about 1% of the total energy spectrum of the sun. Since these rays have quanta of large cross sections of absorption, they are absorbed by mono-atomic and diatomic particles which prevail in the upper atmosphere. These regions of solar energy spectrum are responsible for the dissociation and ionization of the upper atmosphere.

23. Upper Atmosphere Temperature

Figure 23.2 shows the distribution of temperature. (R. Jastrow, Planetary Atmospheres, Proc. Int. School of Physics, Enrico Fermi, Course XXIV edited by B. Rossi, Academic Press, New York, 1964). The temperature profile, up to 85 km is based on balloon and rocket measurement. Above 85 km, the temperature is obtained from rocket and satellite observations.

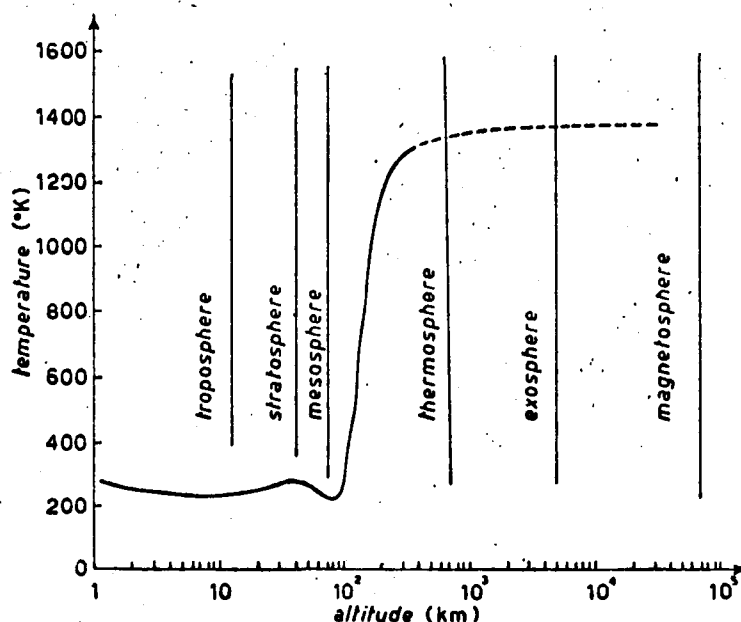


Fig. 23.1-The temperature profile of the Earth's atmosphere

As a result of the absorption by ozone (O_3) of solar ultra-violet radiation in the wave length range 2000A - 3000A, the temperature rises. This explains in Figure 23.1 the temperature increase from the sea level till 50 km height. At altitudes between 50 km and 90 km, the temperature decreases to 200° because of emission of infrared by carbon dioxide (CO_2) and diatomic oxygen (O_2). At higher altitudes the temperature rises again and at 300 km reaches the level of about 1200°k. This rapid temperature increase is caused by photodissociation and by photoionization of oxygen and nitrogen. These processes are produced by solar far-ultraviolet radiation.

Table 23.1 below gives a summary of the effects of solar radiation on the upper atmosphere gases.

SUMMARY OF EFFECTS OF SOLAR RADIATION ON UPPER ATMOSPHERIC GASES

Spectral region (μ)	Reaction	Height (km)	Remarks
0.300-0.210 (Hartley absorption bands)	$O_3 + h\nu \rightarrow O_3 + O^*$ (excited)	50-60	Strong absorption by ozone. Although the reaction takes place for absorbed radiation with wavelengths $< 1.1340 \mu$, ozone absorbs strongly only from 0.300 to 0.210 μ .
0.1925-0.1760 (Runge Schumann absorption bands)	$O_3 + h\nu \rightarrow O_3^*$ (excited) $\rightarrow 2O$ $O_3^* + O_3 \rightarrow O_3 + O$ $O + O_2 + M \rightarrow O_3 + M$	50-80	Comparatively weak absorption. Sequence of ozone formation.
0.1751-0.1200 (Runge Schumann continuum)	$O_3 + h\nu \rightarrow O + O^*$ (excited)	80-110	Strong absorption. Dissociation of O_3 .
0.12157 (Lyman α)	$NO + h\nu \rightarrow NO^+ + e$	60-90	Formation of D region?
0.10247 (Lyman β)	$O_2 + h\nu \rightarrow O_2^+ + e$	90	Contribution to base of E region.
0.1012-0.0910	$O_2 + h\nu \rightarrow O_2^+ + e$	50-80	Weak absorption. Contribution to D region.
0.0910-0.0795	$O + h\nu \rightarrow O^+ + e$	> 200	Very strong absorption. Ionization of O contributes to F_1 and F_2 regions?
0.0795-0.0755	$N_2 + h\nu \rightarrow N_2^+ + e$	140-160	Comparatively weak absorption. Contribution to E_2 region?
0.0744-0.0661	$O_2 + h\nu \rightarrow O_2^*$ (excited) $+ e$	90-120	Strong absorption. Contribution to E_1 region?
0.0661-0.0585	$N_2 + h\nu \rightarrow N_2^*$ (excited) $+ e$	200	Very strong absorption. Contribution to F_1 region.
2×10^{-2} - 1.5×10^{-3} (X-rays)	General ionization	90-450	Contribution to E and F regions.
$\sim 2.5 \times 10^{-4}$	General ionization	60-90	Contribution to D region?

* Based on S. K. Mitra, *Compendium Meteorol.*, p. 245 (1951), and H. Friedman, in *Physics of the Upper Atmosphere* (J. A. Ratcliffe, ed.), p. 133. Academic Press, New York, 1960.

Table 23.1

24. Density Distribution of the Heterosphere

The heterosphere is the region of the upper atmosphere between 200 km and 1500 km. M. Nicolet (Smithsonian Contributions to Astrophysics, Vol. 6, p. 175, 1963) developed a method for calculating the density of the atmosphere by using temperature as the essential parameter. In this section his method is described.

The density, ρ , is defined as a function of the height, z , density, at sea level, and H_ρ = parameter of the vertical distribution of density.

$$\rho = \rho_0 \exp(-z/H_\rho), \quad (24.1)$$

Using the conditions of a perfect gas and of a hydrostatic distribution,

$$\frac{dp}{p} = \frac{dn}{n} + \frac{dT}{T} = -\frac{dz}{H}, \quad (24.2)$$

where p = the total pressure; n = total molecular concentration; T = the absolute temperature and H = the local atmospheric height defined as

$$H = \frac{kT}{mg}, \quad (24.3)$$

where k = Boltzmann's constant; m = the mean molecular mass. The differentiation of (24.3) with respect to H , m , T and g gives the relationships between the respective variations.

$$\frac{dH}{H} = \frac{dT}{T} - \frac{dm}{m} - \frac{dg}{g} \quad (24.4)$$

Introducing the gradient of atmospheric scale height

$$\beta = dH/dz, \quad (24.5)$$

From (24.2) and (24.4) (24.6)

$$\frac{1}{p} \frac{dp}{dz} = -\frac{1}{H},$$

and

$$\frac{1}{\rho} \frac{d\rho}{dz} = -\frac{1+\beta}{H}. \quad (24.7)$$

Since by the definition of the density parameter

$$\frac{1}{\rho} \frac{d\rho}{dz} = -\frac{1}{H_p}, \quad (24.8)$$

the local atmospheric scale height H and the density parameter H_p are related by

$$H = (1+\beta)H_p. \quad (24.9)$$

The equation (24.6) and (24.7) are written as

$$\frac{dp}{p} = -\frac{1}{\beta} \frac{dH}{H}, \quad (24.10)$$

and

$$\frac{d\rho g}{\rho g} = -\frac{1+\beta}{\beta} \frac{dH}{H}, \quad (24.11)$$

The integration of these equations and expansion in series yield

$$\frac{p}{p_0} = \exp \left\{ -\frac{z}{\frac{1}{2}(H+H_0)} \left[1 + \frac{1}{3} \left(\frac{H-H_0}{H+H_0} \right)^2 + \frac{1}{5} \left(\frac{H-H_0}{H+H_0} \right)^4 + \dots \right] \right\} \quad (24.12)$$

and

$$\frac{\rho g}{\rho_0 g_0} = \exp \left\{ -\frac{(1+\beta)z}{\frac{1}{2}(H+H_0)} \left[1 + \frac{1}{3} \left(\frac{H-H_0}{H+H_0} \right)^2 + \frac{1}{5} \left(\frac{H-H_0}{H+H_0} \right)^4 + \dots \right] \right\} \quad (24.13)$$

For conditions where a height interval is less than one scale height, the first approximation can be made-- neglecting terms less than 0.01

$$\frac{\rho g}{\rho_0 g_0} = \exp \left(-\frac{(1+\beta)z}{\frac{1}{2}(H+H_0)} \right), \quad (24.14)$$

For a constant gradient of the scale height from (24.9)

$$\beta = \frac{\beta}{1+\beta}, \quad (24.15)$$

and (24.11) can be written as

$$\frac{d\rho g}{\rho g} = -\frac{1}{\beta} \frac{dH}{H}, \quad (24.16)$$

Equation (24.16) can be redefined with the same order of approximation and after integration

$$\frac{\rho g}{\rho_0 g_0} = \exp \left(-\frac{z}{\frac{1}{2}(H_0+H_{\infty})} \right). \quad (24.17)$$

It is possible to work out a variety of atmospheric models based on various combinations of the scale-height gradients, variations of temperature and mean molecular mass.

The computations lead to the conclusion that the atmospheric conditions are defined by a diffusion distribution and are related to the time of conduction. The temperature appears to be the most important factor in the atmospheric density profile. By adjusting the temperature and its vertical gradient, one can account for variation of density due to diurnal variations, for variations of ultraviolet radiation and for magnetic storm effects.

Extensive studies of satellite orbit perturbations provided insight into the mechanism of atmospheric temperature and density variations.

Five principal density variations were established. They are:

- (1) Day to night variations
- (2) Variations with solar activity
- (3) Variations with geomagnetic activity
- (4) Semi-annual variations
- (5) Latitude dependent seasonal variations

Source: L. G. Jacchia, Special Report No. 184, SAO, 1965. Various efforts were made to evolve a mathematical model, which would quantitatively correlate the density with the parameters indicated above. Considerable progress has been made in this direction, however, we still lack explanations for many phenomena.

It is apparent that solar EUV radiations and solar activities in general, exercise an overriding control over atmospheric density. These phenomena directly affect atmospheric gases and indirectly influence the atmosphere by perturbations of the geomagnetic field. Improvement of our knowledge about the solar-atmospheric interrelation has been significant, however, there is still need for a more complete understanding of the processes involved. Many questions

need asking. In this connection the following questions formulated by

L. G. Jacchia are quoted from the SAO publication (p.14) mentioned above.

"...does it (solar EUV) heat through direct absorption or also through production of ions which are then driven by the earth's magnetic field?

Why are the small variations of the earth's magnetic field accompanied by substantial heating effects?

Is there a permanent corpuscular heat source in the ionosphere?

What causes the semiannual variation?

How much of the seasonal variations is originated in the homosphere and how much is added to them in the thermosphere?"

Hopefully, integrated theoretical studies comprising solar activities and EUV radiation, behavior of atmospheric gases and geomagnetic phenomena, will provide a better understanding of atmospheric processes. These studies will be particularly effective if combined with experimental programs extended over the entire eleven-year solar cycle.

BIBLIOGRAPHY

- ALFVEN, H., Cosmical Electrodynamics, Oxford University Press, London and New York, 1950
- BAESCHLIN, C.F., Geodaesie, Orell Fussli, Zurich, 1948
- BOMFORD, G., Geodesy, Oxford University Press, London and New York, 1952
- BROVAR, V.V. et Al, The Theory of the Figure of the Earth, FTD, AFSC, 1964
FTD-TT-64-930/1+2
- CHAMBERLAIN, J.W., Physics of the Aurora and Airglow, Academic Press, New York, 1961
- CHANDRASEDHAR, S., Radiative Transfer, Oxford University Press, London and New York, 1950
- CHAPMAN, S. and J. BARTELS, Geomagnetism, Oxford University Press, London and New York, 1949
- FLAEGLE, R.G. and J.A. BUSINGER, An Introduction to Atmospheric Physics, Academic Press, New York, 1963
- GLASSTONE, S. Editor, Sourcebook on The Space Sciences, D. Von Nostrand Co., Inc., Princeton, New Jersey, 1965
- HANDBUCH DER PHYSIK, (J. Bartels, ed.), Vols. 47 and 48, Springer, Berlin, 1956
- Handbook of Geophysics, USAF, Rev. Ed., The MacMillan Co., New York, 1961
- HEISKANEN, W.A. and F.A. VENING MEINESZ, The Gravity and Its Gravity Field, McGraw-Hill, New York, 1958
- JEFFEREYS, Sir H., The Earth, Fourth Ed. Cambridge University Press, Cambridge, 1959
- JOOS, G., Theoretical Physics, Second Ed. (translated by Ira M. Freeman), Hafner, New York, 1950
- KELLOGG, O.D., Foundations of Potential Theory, Dover, New York, 1953
- LE GALLEY, D.P. and A. ROSEN, Editors, Space Physics, John Wiley, New York, 1964
- MACMILLAN, W.D., The Theory of the Potential, Dover, New York, 1958
- MITRA, S.K., The Upper Atmosphere, Asiatic Society, Calcutta, 1952
- PARKER, E.N., Interplanetary Dynamical Processes, Interscience Publishers, New York, 1963
- RATCLIFFE, J.A. Editor, Physics of the Upper Atmosphere, Academic Press, New York, 1960

(cont.)

SEARS, F.W., An Introduction to Thermodynamics, The Kinetic Theory of Gases and Statistical Methods, Addison-Wesley, Reading, Massachusetts, 1953

TVERSKOI, P.N., Physics of the Atmosphere, Israel Program for Scientific Translation, Jerusalem, 1965.

UNSÖLD, A., Physik der Sternatmosphären, Springer, Berlin, 1955

214
N 6 7. 8. 0. 4 6. 5

SPACE SCIENCE - *"SPACE GUIDANCE AND CONTROL"*

IV

by

Nasser Nahi

SECTION I

Introduction to Laplace Transformation and Block Diagram Representation

1. Introduction

The dynamical systems considered in this chapter can be represented by one or more ordinary linear differential equations with constant coefficients. While these equations can, in general, be solved by direct methods, application of the operational methods of Laplace transformation has the following advantages:

- 1) The differentiation and integration operations are substituted by simple algebraic operations and consequently solution of the differential equation is reduced to a problem in algebra. The solution is in the operational form and tables are available to transform these solutions to the time domain. Many of the very important properties of the time domain solution can be directly obtained from its operational form.
- 2) Boundary or initial conditions are automatically included.
- 3) It provides a simple block diagram representation of the system.

2. Laplace Transformation

Let a function of time $f(t)$ be piecewise continuous and of exponential order.* The Laplace transform of $f(t)$ is denoted by $\mathcal{L} f(t)$ or $F(s)$ and is given by

$$\mathcal{L} f(t) = F(s) = \int_0^{\infty} f(t) e^{-st} dt \quad (2.1)$$

The variable s is a complex quantity of the form $\sigma + j\omega$. As an example, the transform of the step function $f(t) = u(t)$ is

$$\mathcal{L} u(t) = \int_0^{\infty} e^{-st} dt = \frac{1}{s} \quad (2.2)$$

*The function $f(t)$ is of exponential order if there exists a constant such that $e^{-\sigma t} |f(t)|$ is bounded for all t larger than some finite T .

The Laplace transform of $f(t) = e^{-\alpha t}$

$$\mathcal{L} e^{-\alpha t} = \int_0^{\infty} e^{-\alpha t} e^{-st} dt = \frac{1}{s+\alpha} \quad (2.3)$$

2.1 Properties of Laplace Transformation

The following relations between $f(t)$ and $F(s)$ can be established by the direct application of Equation (2.1).

a) Linearity

$$\mathcal{L} [a f(t)] = a F(s) \quad (2.4)$$

$$\mathcal{L} [a_1 f_1(t) + a_2 f_2(t)] = a_1 F_1(s) + a_2 F_2(s) \quad (2.5)$$

b) Translation in s domain

$$\mathcal{L} [e^{-at} f(t)] = F(s+a) \quad (2.6)$$

c) Differentiation

$$\mathcal{L} \left[\frac{d}{dt} f(t) \right] = sF(s) - f(0^+) \quad (2.7)$$

$$\mathcal{L} \left[\frac{d^n}{dt^n} f(t) \right] = s^n F(s) - s^{n-1} f(0^+) - \dots - \frac{d^{n-1} f}{dt^{n-1}} (0^+) \quad (2.8)$$

d) Integration

$$\mathcal{L} \left[\int_0^t f(t) dt \right] = \frac{F(s)}{s} + \frac{\int_0^0 f(t) dt}{s} \quad (2.9)$$

e) Final Value

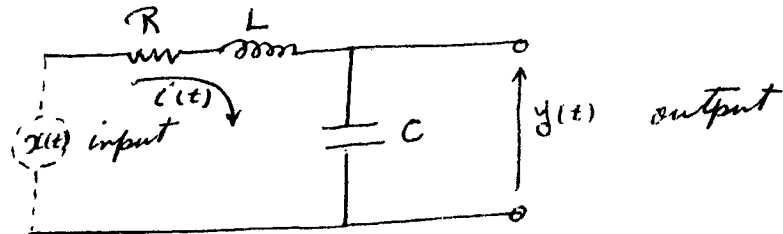
$$\lim_{s \rightarrow 0} sF(s) = \lim_{t \rightarrow \infty} f(t) \quad (2.10)$$

f) Initial value

$$\lim_{s \rightarrow \infty} sF(s) = \lim_{t \rightarrow 0} f(t) \quad (2.11)$$

3. Application of Laplace Transformation to Solution of Differential Equations; Transfer Function.

Let us consider the following network



where $x(t)$ is the value of an input voltage source and $y(t)$ is the output of the network. The following equation can be written

$$x(t) = R i(t) + L \frac{d}{dt} i(t) + \frac{1}{C} \int_0^t i(t) dt \quad (3.1)$$

where

$$\frac{1}{C} \int_0^t i(t) dt = y(t) \quad (3.2)$$

From Eq. (3.2)

$$\frac{d}{dt} y(t) = \frac{1}{C} i(t) \quad (3.3)$$

$$\frac{d^2}{dt^2} y(t) = \frac{1}{C} \frac{d}{dt} i(t) \quad (3.4)$$

Substituting 3.2 - 3.4 into 3.1 yields

$$x(t) = LC \frac{d^2}{dt^2} y(t) + RC \frac{d}{dt} y(t) + y(t) \quad (3.5)$$

3.5 is a differential equation relating $x(t)$ and $y(t)$. Assuming zero initial conditions, i.e. $y(0) = 0$; $\frac{d}{dt} y(t)|_{t=0} = 0$ and taking transform of both sides, we have

$$X(s) = LC s^2 Y(s) + RC s Y(s) + Y(s) \quad (3.6)$$

The transfer function is defined to be the ratio of transform of the output to transform of input and is denoted by $G(s)$.

$$G(s) = \frac{Y(s)}{X(s)} \quad (3.7)$$

From 3.6

$$G(s) = \frac{1}{LCs^2 + RCs + 1} \quad (3.8)$$

Hence

$$Y(s) = G(s) X(s) \quad (3.9)$$

Suppose it is desired to find the response to $x(t) = u(t)$, then

$$Y(s) = \frac{1}{LCs^2 + RCs + 1} \frac{1}{s} \quad (3.10)$$

Which is the solution in operational form. If solution in time domain is required, the function $y(t)$ which yields Eq. (3.10) should be found. This process is referred to as inverse transformation. Standard tables are available which will facilitate the process of inverse transformation.

The representation of the input output relationships by means of Laplace transformation suggest a simple method of describing systems through block diagram notation. As an example, a single-degree-of-freedom stabilized platform is discussed below.

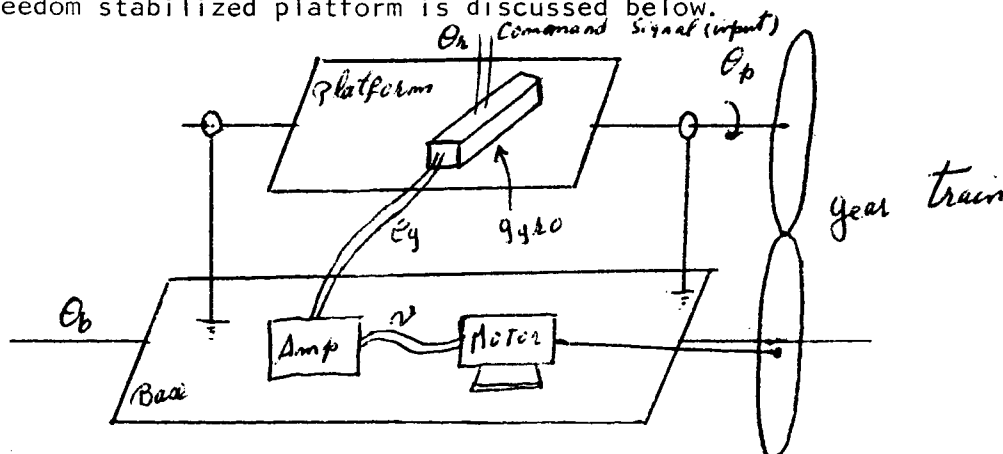


Figure 3.1 - Single-Degree-of-Freedom-Stabilized Platform

The purpose of this control system is to keep the orientation of the platform fixed with respect to a reference for any angular movement of the base represented by θ_b . The following set of differential equations govern the dynamics of the system shown in Fig. 3.1:

- 1) The gyro is an integrating gyro and produces a signal proportion to the difference between the desired platform angle θ_r and the actual angle θ_p

$$e_g = K_1 (\theta_r - \theta_p) \quad (3.11)$$

where K_1 is the constant of proportionality.

The amplifier amplifies the signal e by a factor K_2

$$v = K_2 e_g \quad (3.12)$$

The motor produces a torque f proportional to v (it is assumed that f is unaffected by the motor shaft speed).

$$f = K_3 v \quad (3.13)$$

The torque f through the gear train with a gain K_4 causes the platform to rotate. The platform has an inertia J and damping B

$$K_4 f = J \ddot{\theta}_p + B \dot{\theta}_p \quad (3.14)$$

Taking Laplace transform of equations (3.11) through (3.14) we have

$$E_g(s) = K_1 (\theta_r(s) - \theta_p(s)) \quad (3.15)$$

$$V(s) = K_2 E_g(s) \quad (3.16)$$

$$F(s) = K_3 V(s) \quad (3.17)$$

$$K_4 F(s) = J s^2 \theta_p(s) + B s \theta_p(s) = s \theta_p(s) [J s + B] \quad (3.18)$$

Now the following block diagram represents the set of equations (3.15) through (3.18)

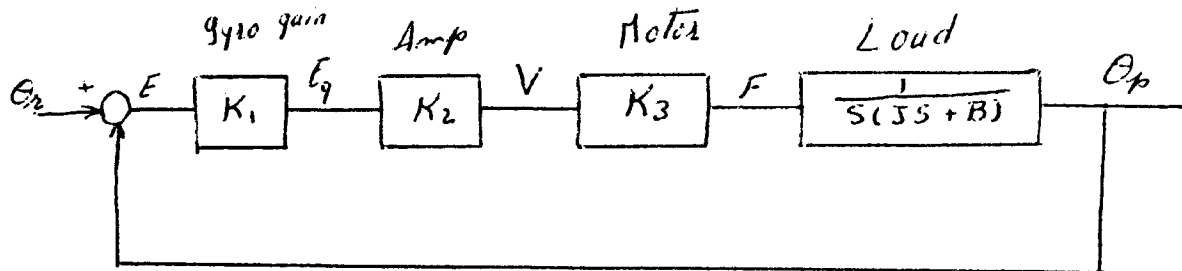


Figure 3.2 - Block Diagram for Stabilized Platform

When $\dot{\theta}_r = \dot{\theta}_p$ then $E = 0$ and consequently $F = 0$ which in turn will cause no change in θ_p . The error E may be different from zero, either by a change in command $\dot{\theta}_r$ or through a disturbance introduced by the motion of the base θ_b .

SECTION II

Feedback Systems and Stability

4. Open-Loop vs. Closed-Loop

The function of stabilizing the platform in the example of the previous section can also be accomplished in the following manner. Referring to Fig. 3.1, if a given change in θ_p is desired, then an appropriate signal V can be applied to the motor to cause the right amount of change in θ_p . The system block diagram is in this case presented by Fig. 4.1

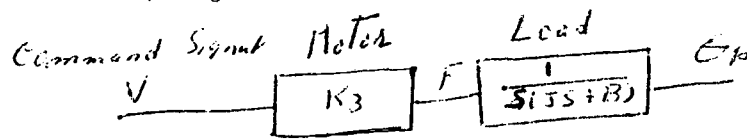
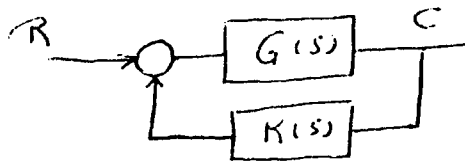


Figure 4.1 - Open-Loop Control of θ_p

This type of control is called open-loop control, and requires precise knowledge of the constants K_3 , J and B and is precise only if the base does not rotate after the initiation of command. In practice the parameters defining the dynamics of controlled objects either are not known with sufficient accuracy or their value may vary due to various conditions of the environment which are very difficult to account for. Consequently open-loop control is very sensitive to parameter variation and completely ineffective for eliminating any effect of disturbance inputs, such as motion of the base. Closed-loop systems, if designed properly, can reduce the effect of these problems to a point where the design is acceptable.

The following relationship between the transfer function of an open-loop system $G(s)$ and that of a corresponding closed-loop system $H(s)$ with a feedback transfer function $K(s)$ is derived.



$$\frac{C(s)}{E(s)} = G(s) \quad (4.1)$$

$$\frac{C(s)}{R(s)} = H(s) \quad (4.2)$$

$$C(s) = E(s)G(s) = [R(s) - K(s)C(s)]G(s) \quad (4.3)$$

and finally, from (4.2) and (4.3)

$$H(s) = \frac{G(s)}{1 + K(s)G(s)} \quad (4.4)$$

5. Stability

A system is called stable if it returns to the "rest" condition after an initial perturbation. It is evident that this property is required of all practical control systems. In the stabilized platform example of Section 3, it means that, for example, in the absence of any change in command signal and base movement, if the platform is perturbed from its rest position, it returns to that condition automatically. Consequently, it is important to determine the stability condition of a system from the governing equations.

A differential equation describing a stable linear system has the property that its characteristic equation has roots with negative real parts. This guarantees that the homogeneous solution will have an exponentially decaying element ($e^{-\alpha t}$, $\alpha > 0$) in each additive part, hence, all transients will decay out with time. The characteristic equation of a system is readily available as the denominator polynomial of the closed-loop transfer function. The stabilized platform example of Section 3 will be used here for illustration of stability analysis.

From block diagram of Fig. 3.2, the closed-loop transfer function is given by

$$\frac{\Theta_p(s)}{\Theta_a(s)} = \frac{K}{s(s^2 + B)s + K} = \frac{\frac{K}{s(s+B)}}{1 + \frac{K}{s(s+B)}} \quad (5.1)$$

$$K = K_1 K_2 K_3$$

Hence

$$\frac{\Theta_p(s)}{\Theta_r(s)} = \frac{K}{Js^2 + Bs + K} \quad (5.2)$$

The differential equation relating input $\theta_r(t)$ to the output $\theta_p(t)$ can be deduced from (5.2) directly by identifying S as d/dt (the differential operator)

$$J \ddot{\theta}_p(t) + B \dot{\theta}_p(t) + K \theta_p(t) = K \theta_r(t) \quad (5.3)$$

which has the characteristic equation given by the denominator of (5.2)

$$J s^2 + B s + K = 0 \quad (5.4)$$

The polynomial (5.4) has the following two roots

$$-\frac{B}{2} \pm \sqrt{\frac{B^2}{4} - KJ} \quad (5.5)$$

Since B , K and J are positive quantities, these two roots both have negative real parts for all values of these parameters. Notice that if feedback was positive instead of negative (i.e. for example, when the gear train is connected wrong) then (5.1) should be replaced by

$$\frac{\frac{K}{s(Js+B)}}{1 - \frac{K}{s(Js+B)}} \quad (5.6)$$

which yields the following two roots for the characteristic equations

$$-\frac{B}{2} \pm \sqrt{\frac{B^2}{4} + KJ} \quad (5.7)$$

which clearly indicates that one of the roots is always a positive number.

In the area of classical control systems, a number of graphical and numerical techniques are available by which the stability of a linear system can be determined.

SECTION III
Space Vehicle Systems

6. General Description of Mathematical Models for Various Space Vehicle Operations

The control functions of a space vehicle may be divided into three parts a) attitude control b) guidance and c) tracking. In the following the above operations are described in brief.

a) Attitude Control - There are various reasons for the requirement that the orientation of a vehicle remain at a desired relationship, with respect to some reference set of axis. To name a few, a known orientation is needed when the vehicle is thrust, a fixed orientation is desired when there are astronauts on board, a controlled orientation is necessary when an on-board camera (or any similar instrument) is to point in a specific direction. The reference system may be a fixed set of directions in space, which is obtained by means of an inertial platform or by direction of location of some stars. In case of an earth satellite, a reference system may have one axis which at any time passes through the center of the earth.

b) Guidance of Space Vehicles - Any space vehicle probably has a mission to perform, and that requires the transfer of the vehicle from one point in space to another. The requirement and conditions of any mission in general, adds more restrictions on how the transfer is to be accomplished and what trajectory to be followed. For example, it may be desired to land a vehicle on moon. A specific area on the surface of the moon may be chosen as the desired landing position. In order not to crash land, restrictions should be put on the terminal velocities of the vehicle. Furthermore, the value of the thrusts of the on-board engines are limited by a known quantity.

c) Tracking - Tracking is an important function to be carried out on board many future space vehicles. In case of rendezvous and docking of two vehicles, the required accuracies usually exceed those obtained by means of ground tracking of both vehicles. Future vehicles will definitely be required to be more autonomous (self governing) than those of the past. This appears in form of lessening necessity for reliance

on information supplied by a ground base. Consequently, the necessary information for various pursuit and rendezvous missions have to be obtained on board by means of radar, optical or passive detection.

The above three functions will be treated in more detail in the next three sections. In the remainder of this section, some general properties of the mathematical models for the vehicle in the various modes of operation will be discussed.

In any of the desired operations, there is a part or all of the vehicle with dynamics which have to be controlled and there is mechanism for inducing the necessary control. In Fig. 6.1 these are referred to as controlled object and the controller, respectively. The information available to the controller is the reference (or command) input or inputs and the output variable or variables which are fed back through the use of various sensors.

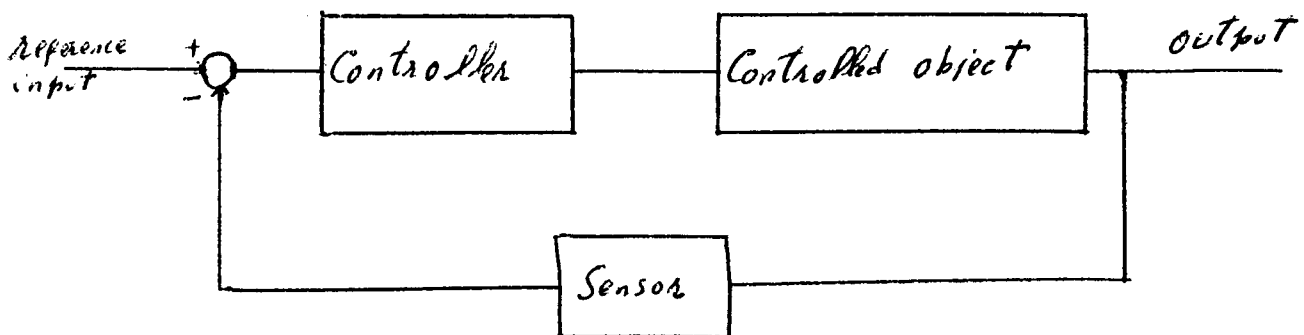


Figure 6.1 - General Block Diagram of a Control System.

For example in the case of tracking the controlled object is the antenna system, the controller is the drive (motor) which can re-orient the antenna and the sensor, may be a rate integrating gyroscope indicating the deviation of the center beam of the antenna from a reference axis. In the case of attitude control, the controlled object is identified by a set of equations representing the dynamics of the body of the vehicle,

including the antenna system and any other object occupying the vehicle, with respect to a coordinate system. The controller may be a set of small reaction jets which when fired, will exert a torque causing a body reorientation. The sensors are a number of position and rate gyroscopes yielding the angular and angular rate deviations of the body with respect to a set of reference axis. Finally, in the case of guidance, the whole system including the antenna system and all the other instruments and astronauts can be assumed as a point mass which will represent the controlled object. The controller is one or more propulsion mechanisms which exert a vector valued force on the point mass (the controlled object). The sensors may be the tracking system on board which in turn will supply information concerning any deviation of the vehicle position, velocity and acceleration with respect to some desired values.

From the above discussion, it is seen that, for the same space vehicle, the controlled object may involve a part or whole of the vehicle depending on various functions to be performed.

7. Attitude Control of Space Vehicle Systems

The attitude control of a vehicle, i.e. maintaining a desired orientation for the body of the vehicle, is basically accomplished by two different procedures, passive and active. Passive attitude control is usually used for unmanned vehicles where a relatively sophisticated control of the orientation may not be necessary. The object here is to design a system which is stable in the vicinity of some desired orientation in other words, any perturbation about the desired orientation reduce to zero in time. The main sources of perturbing torques are: gravity gradient; atmospheric pressure; electromagnetic induction and solar radiation. The methods of passive stabilization include: spin stabilization; balancing one perturbing torque against another; energy dissipation and tuned pendulum. The active attitude control is accomplished through the use of applied torques (e.g. generated by jets mounted on the body of the vehicle) in order to perform desired corrections in the orientation of the vehicle. Active attitude control is necessary when the mission calls for high accuracy and speed in response to a command input.

Equations of Motion

Let Fig. 7.1 represent a space vehicle. The point 0 is the center of mass, ox, oy, oz are a set of orthogonal axis fixed on the body.

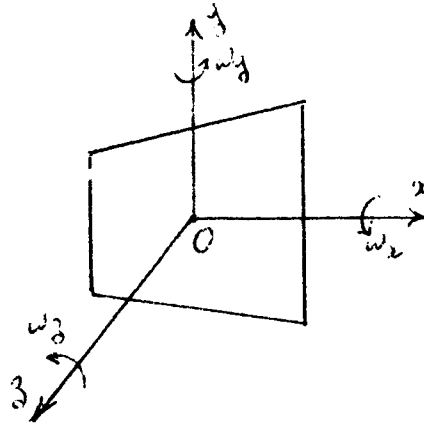


Figure 7.1 - Space Vehicle

The angular velocities with respect to x, y, z axis are denoted by $\omega_x, \omega_y, \omega_z$. The moment of inertias are referred to as I_x, I_y, I_z . There may be one swiveling engine which induces a torque T with components T_x, T_y, T_z or a set of fixed, body mounted thrusters may produce these component values. The Newton's law of motion gives the following relationship.

$$\text{Applied torque } T = \text{Rate of change of angular momentum with respect to an inertial coordinate system.} \quad (7.1)$$

With respect to the body fixed axis x, y, z the law of motion assumes the well known Euler equations.

$$T_x = I_x \dot{\omega}_x + (I_z - I_y) \omega_y \omega_z \quad (7.2)$$

$$T_y = I_y \dot{\omega}_y + (I_x - I_z) \omega_z \omega_x \quad (7.3)$$

$$T_z = I_z \dot{\omega}_z + (I_y - I_x) \omega_x \omega_y \quad (7.4)$$

For a body with three axis of symmetry $I_x = I_y = I_z = I$ and consequently (7.2) to (7.4) become

$$T_x = I \dot{\omega}_x \quad (7.5)$$

$$T_y = I \dot{\omega}_y \quad (7.6)$$

$$T_z = I \dot{\omega}_z \quad (7.7)$$

In general, even when the above symmetry does not exist, the equations (7.2) to (7.4) can be approximated by (7.5) to (7.7) since $\omega_x, \omega_y, \omega_z$ are kept relatively small.

Since equations (7.5) to (7.7) are completely independent of each other, only one such as the rotation about the x axis, is considered in the following. Let θ_x be the rotation angle about x axis with respect to a reference and θ_{xd} be the desired value of the rotation about this axis. We have the following equations

$$T_x = I \dot{\omega}_x \quad (7.8)$$

$$\omega_x = \dot{\theta}_x \quad (7.9)$$

Furthermore, the thruster produces a torque proportion to the difference $\theta_{xd} - \theta_x$ with a constant of proportionality K . Therefore the following block diagram results.

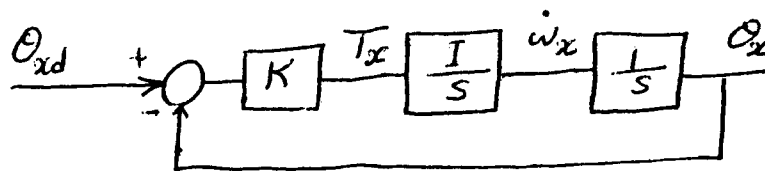


Figure 7.2 - Block Diagram

The difference quantity $\theta_{xd} - \theta_x$ is the output of a rate integrating gyro with a reference axis θ_{xd} . The output Laplace transform is

$$\theta_x(s) = \theta_{xd}(s) \frac{\frac{KI}{s^2}}{1 + \frac{KI}{s^2}} = \theta_{xd}(s) \frac{KI}{s^2 + KI} \quad (7.10)$$

This represents an unstable system since the roots of the denominator polynomial do not have negative real parts. To see this better, let us assume we would like to turn the vehicle around the x axis by 1 *unit*. Then

$$\theta_x(s) = \frac{1}{s} \frac{KI}{s^2 + KI} = \frac{1}{s} - \frac{s}{s^2 + KI} \quad (7.11)$$

Consequently

$$\theta_x(t) = 1 - \cos \sqrt{KI} t \quad (7.12)$$

which clearly shows that $\theta_x(t)$ rather than approaching 1 will oscillate about the desired value. What is now needed to achieve the desired goal is referred to as a compensator. A very common way of stabilizing the above system is to add a rate feedback path. A rate gyroscope will produce an output proportion to angular rate $\dot{\theta}_x$. Let the constant proportionality be K_1 . The block diagram of Fig. 2 is modified as

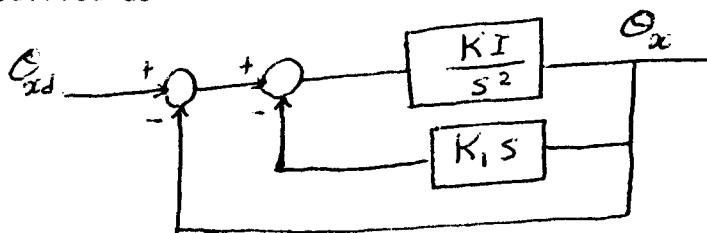


Figure 7.3 - The Compensated System

The transform of the output is

$$\theta_x(s) = \theta_{xd}(s) \frac{\frac{KI}{s^2}}{1 + \frac{KI(1 + K_1 s)}{s^2}} = \theta_{xd}(s) \frac{KI}{s^2 + KK_1 s + KI} \quad (7.13)$$

The system is now stable (the roots having negative real parts) and the solution to $\theta_{xd}(t) = 1$ is (for K_1 small)

$$\theta_x(t) = 1 + \alpha e^{-\frac{KK_1 I}{2} t} \sin(\omega_c t + \varphi)$$

$$\alpha = \frac{1}{\sqrt{1 - \frac{1}{4} K_1^2 K I}}$$

$$\omega_c = \sqrt{K I} \sqrt{1 - \frac{1}{4} K_1^2 K I}$$

$$\varphi = \tan^{-1} \frac{\sqrt{1 - \frac{1}{4} K_1^2 K I}}{\frac{1}{2} K_1 \sqrt{K I}}$$
(7.14)

which clearly shows that as time is increased, the desired orientation will be approached.

In the above analysis, the operation of the thrusters were assumed linear, i.e., the output torque T_x was assumed to be proportional to its input, $\theta_{xd} - \theta_x$. Since reaction jets are commonly used in space vehicle applications (due to reliability and weight considerations) the operation is far from linear and is approximately presented by Fig. 7.4 where e is the input to the torque.

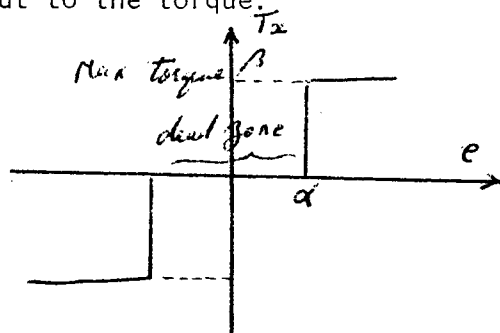


Figure 7.4 - The Characteristic of Torquers

The block diagram of Fig. 7.3 should be modified.

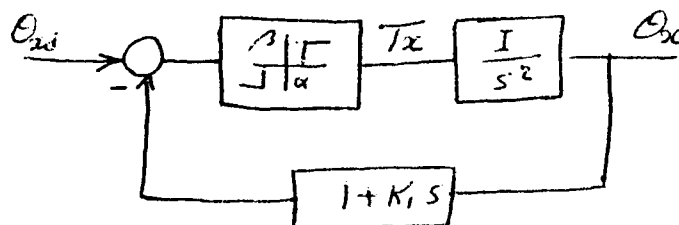


Figure 7.5 - Nonlinear Attitude Control Loop

If $K_1 = 0$ (i.e. no rate feedback) it is easy to show that the output will oscillate about the desired input θ_{xd} (e.g. for $\theta_{xd} = \text{constant}$). As an example, let the desired rotation θ_{xd} be unity and this is applied to the system as a step command. If $\alpha > 1$ (α is the width of the deadzone or inactive zone of the torquers) the system will not become active. If $\alpha < 1$ then a torque $T_x = \beta$ will be applied to the vehicle which will yield ($e(t)$ is the input to the nonlinearity)

$$\theta_x(t) = \frac{1}{2} I \beta t^2 \quad (7.15)$$

$$e(t) = 1 - \frac{1}{2} I \beta t^2 - K_1 I \beta t \quad (7.16)$$

This continues until $e(t)$ is reduced to $+\alpha$ when the body will keep moving then on its own inertia until $e(t) = -\alpha$. At this time, a reverse torque will be applied. Time plot of this operation with and without K_1 term (i.e. $K_1 = \text{nonzero}$ and $k_1 = 0$) will reveal that the system has a damped response for $K_1 > 0$ and oscillatory response for $K_1 = 0$. In order to verify this assertion, let us define an equivalent gain for the nonlinearity. Assume the input to the nonlinearity is a sine wave $E \sin \omega t$. The output waveform is sketched in Fig. 7.6.

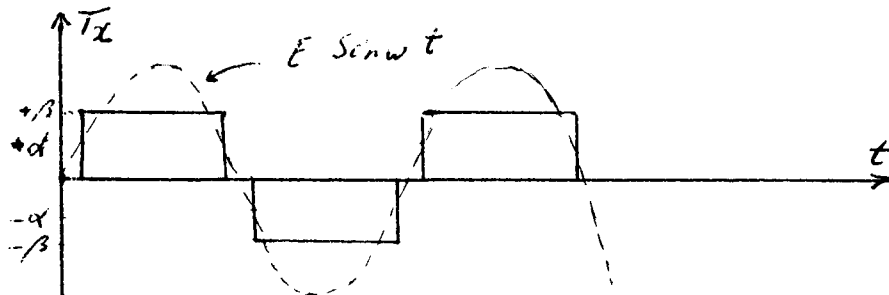


Figure 7.6 - Response of Torquer to Sine Input

The fundamental component of the output waveform is obtained by a fourier series expansion.

$$\text{Magnitude of Fundamental Comp. of } T_2 = \frac{4\beta}{\pi} \sqrt{1 - \frac{\alpha^2}{E^2}} \quad E > \alpha \quad (7.17)$$

$$= 0 \quad E < \alpha$$

The ratio $\frac{4\beta}{\pi} \frac{\sqrt{1 - \frac{\alpha^2}{E^2}}}{E}$ is referred to as the describing function of the nonlinearity.

The block diagram of Fig. 7.5 can be represented as

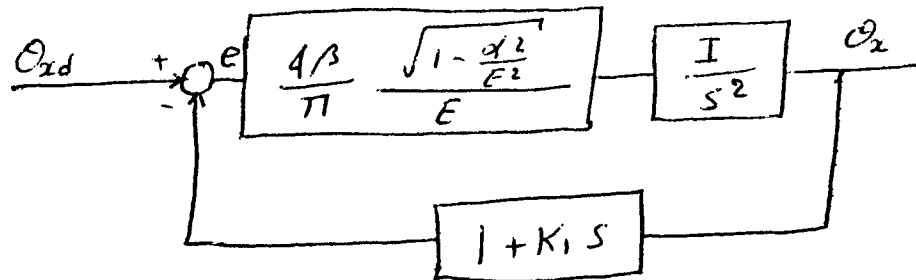


Figure 7.7 - Block Diagram with Describing Function

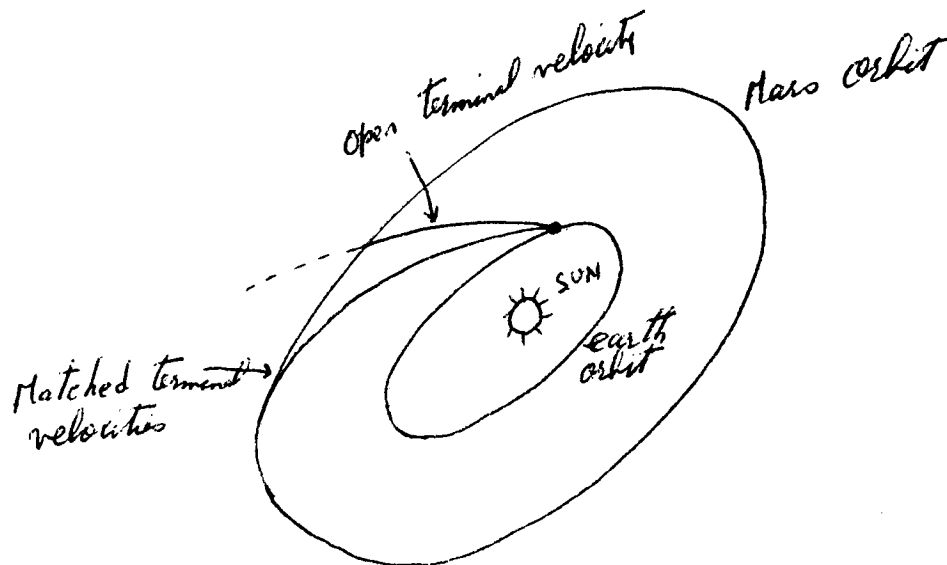
Now let us conjecture that the loop is not stable for some positive value K_1 . Therefore θ_x will oscillate and there is a sinusoidal value for $e(t)$ such as $e = E \sin st$ but with any value for E the describing function is a simple gain and we have already shown that for any gain, the system cannot have any sustained oscillation for $K_1 \neq 0$. Consequently, the conjecture is disproved. It can easily be seen that the conjecture is true if $K_1 = 0$.

8. Guidance of Space Vehicles

The transfer of a vehicle from a point in space and a set of initial values for velocities and accelerations to a desired location satisfying certain conditions on final velocities, etc. is the function of the guidance of the vehicle. It can be accomplished in two ways: a) open-loop and b) closed-loop.

a) Open Loop Guidance

In this scheme, a trajectory which the vehicle should follow in space is determined a priori. This may be a trajectory satisfying the desired initial and terminal conditions or, in addition, may be an optimal trajectory in some sense. For example, it may accomplish the mission with minimum amount of fuel expenditure, or in minimum time. In cases where the final conditions cannot be achieved exactly, the trajectory may result in minimum value for some function of the terminal miss distance. The trajectory is a function of the mass of the vehicle, maximum value of available torque, initial and terminal conditions and the nature of desired optimality. For example, in case of solar sailing when a vehicle is leaving Earth orbit and reaches Mars orbit in minimum time, two optimum trajectories result depending on whether the terminal velocities of the vehicle should match that of the orbit or they can assume any arbitrary value. Obviously, when the terminal velocities are to be matched, the optimum time of transfer is longer since this matching is an additional constraint.



8.1 - Solar Sailing Example

When, by some method, a desired trajectory has been determined, with respect to an inertial coordinate system, in order to maintain the vehicle on that trajectory, it is required to determine the instantaneous position of the vehicle with respect to that coordinate system. This can be accomplished by two means: 1) use the rate gyro's and accelerometers on the vehicle ~~and~~ to compute the trajectory which the vehicle is following, 2) to track the vehicle from another position whose movement with respect to the coordinate system is known.

In case where, at some time, the vehicle trajectory and desired trajectory do not coincide, it is necessary to make trajectory corrections by using the on-board engines. A swiveling engine or body fixed engine can be used to apply thrust (torque) in appropriate direction, the difference being that in the latter case the vehicle orientation is determined by the direction of required correction.

b) Closed Loop Guidance

In many missions, the requirements on the accuracy of achieving the terminal conditions is such that an open loop guidance is not satisfactory. This may be because of the movement of the destination point (target) or errors in determining the vehicle trajectory, or difficulties in making exact trajectory correction. Furthermore, it is always desired to make vehicles more autonomous (self governing). Consequently, in many cases, the open loop guidance either is not used at all, or it is only used during the midcourse, that is, transferring the vehicle from the initial conditions to some vicinity of the destination and then performing the last portion of guidance in "closed loop" fashion. A number of examples of closed loop guidance is given in the following:

1) Pursuit

When an intercepting vehicle with initial velocity vector V_I tries to intercept with a target vehicle, with say, a constant velocity V_T , one possibility is that the interceptor orients its velocity vector such that it always goes through the instantaneous position of the target. This is the example of the dog chasing a rabbit. As a two

dimensional example, let $\mathcal{X}(t)$ be the target path (straight line) with the initial conditions $\mathcal{X}(0) = 0$, $\dot{\mathcal{X}}(0) = v_T(\mathcal{X}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix})$. Let the interceptor be at $y = y_0$ at $t = 0$. The trajectory that the interceptor will follow is sketched in Fig. 8.2.

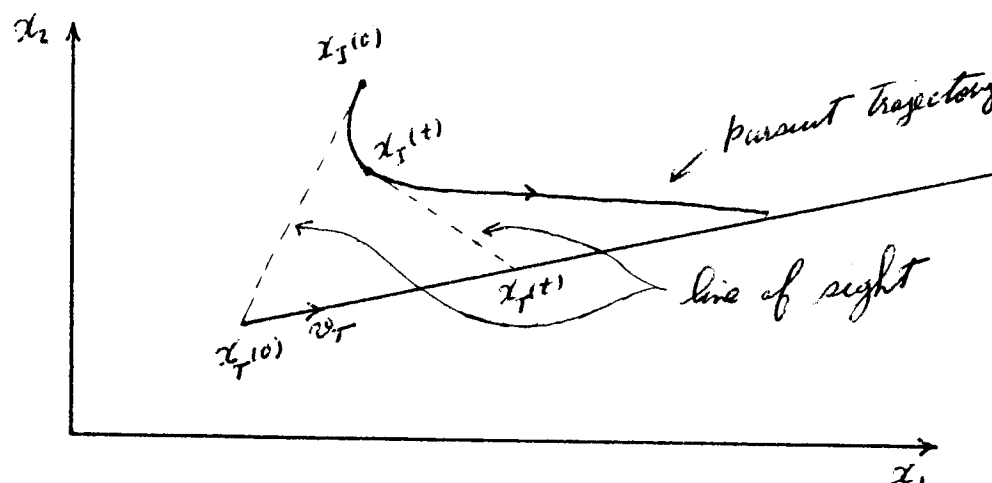


Fig. 8.2 - Pursuit Navigation

It is clearly seen that, although the target has a very simple flight path, the interceptor has to go through a reasonably complicated maneuver. This is mainly due to the inefficiency of pursuit navigation. The interceptor has to continuously apply thrust perpendicular to instantaneous velocity vector in order to keep the velocity vector coincident with the line-of-sight (line-of-sight is the vector from instantaneous interceptor to target position, this is abbreviated by LOS and is given with respect to an inertial coordinate system). A change in direction of the velocity vector can be obtained by applying acceleration in a direction perpendicular to the velocity vector which is accomplished by either orienting the swiveling engine or the whole vehicle in case of body fixed jets in the appropriate direction. A mathematical model for this system is

derived in the following. Fig. 8.3 represents a schematic diagram of the space vehicle target system. The interceptor should attempt to maintain the orientation of its velocity vector in the appropriate direction. A change in the orientation of V_I is accomplished by applying force perpendicular to the direction of V_I . This force will produce a rate of change of the angle δ , i.e. $\dot{\delta}$ is proportion to the applied thrust.

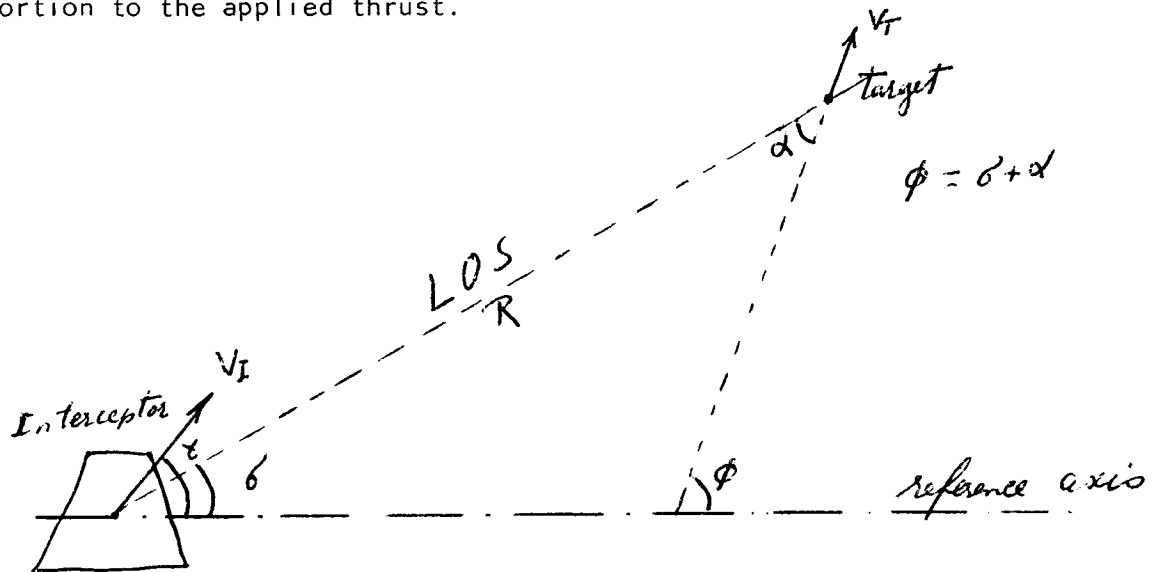


Fig. 8.3 - Schematic Diagram for Interceptor-Target System

Let

- V_I interceptor velocity (constant in magnitude)
- V_T target velocity (constant in magnitude)
- a_{Tn} the component of target acceleration perpendicular to LOS
- R range

Let us first assume $a_{Tn} = 0$. We have the following equations

$$R \dot{\delta} = V_T \sin(\phi - \delta) - V_I \sin(\delta - \delta) \quad (8.1)$$

$$\dot{R} = V_T \cos(\phi - \delta) - V_I \cos(\delta - \delta) \quad (8.2)$$

Differentiating, 8.1 yields

$$\dot{R} \dot{\epsilon} + R \ddot{\epsilon} = -\dot{\epsilon} V_T C_0 (\gamma - \epsilon) - \dot{\gamma} V_I C_0 (\gamma - \epsilon) + \dot{\epsilon} V_I C_0 (\gamma - \epsilon) \quad (8.3)$$

Substituting 8.2 into the right hand side of 8.3 results

$$\dot{R} \dot{\epsilon} + R \ddot{\epsilon} = -\dot{\epsilon} \dot{R} - \dot{\gamma} V_{IR} \quad (8.4)$$

$$\text{where } V_{IR} = V_I C_0 (\gamma - \epsilon)$$

= Component of interceptor velocity in the direction of LOS.

From 8.4

$$\ddot{\epsilon} = -\frac{2\dot{R}}{R} \dot{\epsilon} - \frac{V_{IR}}{R} \dot{\gamma} \quad (8.5)$$

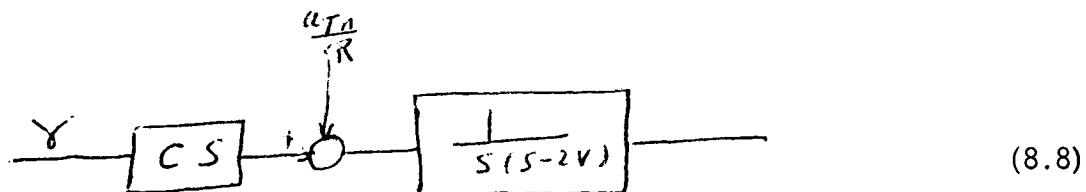
If a_{Tn} was not zero, it would simply add to the $\ddot{\epsilon}$ a quantity equal to a_{Tn}/R , hence

$$\ddot{\epsilon} = -\frac{2\dot{R}}{R} \dot{\epsilon} - \frac{V_{IR}}{R} \dot{\gamma} + \frac{a_{Tn}}{R} \quad (8.6)$$

In order to develop a block diagram, let us assume that over small ranges in time, $\dot{R}/R = -V = \text{Constant}$ and V_{IR} and R are also constant. Taking the Laplace transform of (8.6) yields ($V_{IR}/R = C$)

$$(s^2 - 2Vs) \epsilon = -Cs\gamma + \mathcal{L} \frac{a_{Tn}}{R} \quad (8.7)$$

This yields the following block diagram



The loop is closed through the navigation law which in this case attempts to line up V_I with LOS. This can be accomplished by varying γ proportion to $\gamma - \delta$ and in a direction which reduces the magnitude of $\gamma - \delta$. For example

$$\dot{\gamma} = -\lambda(\gamma - \delta) \quad (8.9)$$

λ is a constant called navigation constant. From (8.9)

$$\gamma(s+1) = \lambda \mathcal{L} \delta \quad (8.10)$$

Finally the following block diagram is obtained

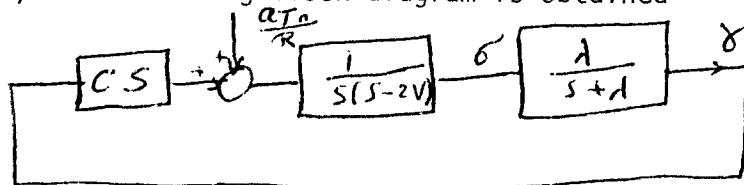


Fig. 8.4 - Block Diagram for Pursuit Navigation (8.11)

It is clearly seen that since the loop tries to make $\delta = \gamma$ and since δ is varying during the flight, consequently the interceptor has to apply acceleration all during the flight.

2) Proportional Navigation.

The proportional navigation is the guidance policy which keeps the LOS non-rotating. Let the LOS angle with respect to some fixed coordinated system be δ and its derivative (LOS rotation)

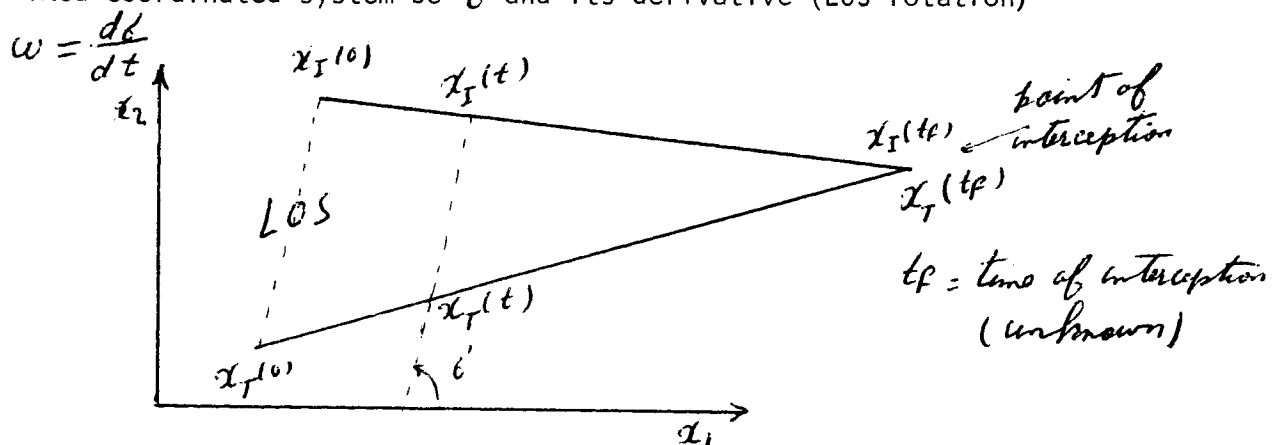


Fig. 8.5 - Proportional Navigation

From Fig. 8.5, it is evident that if LOS is non-rotating, the interception will occur at some time $t_f > 0$. The closed loop system is mechanized such that it "nulls" out any existing LOS rate. The information about the LOS rate is obtained by the tracking system. The equations of motion and a block diagram for the system is derived in the following.

The derivation of Eq. (8.6) is applicable here. The difference is only in the navigation law. Here the control loop tries to zero out any rotation of LOS, namely $\dot{\sigma}$. This is done by letting

$$\ddot{\sigma} = -\lambda \dot{\sigma} \quad (8.12)$$

Substituting (8.12) in (8.6) yields

$$\ddot{\sigma} = -\frac{2\dot{R}}{R} \dot{\sigma} + \frac{\lambda V_{IR}}{R} \dot{\sigma} + \frac{a_{Tn}}{R} \quad (8.13)$$

Which yields the following block diagram

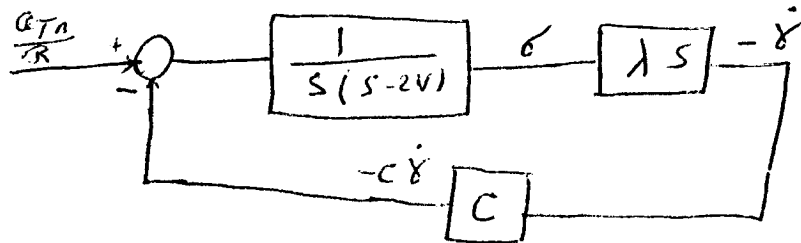


Fig. 8.6 - Block Diagram for Proportional Navigation

The transfer function between a_{Tn}/R and $\ddot{\sigma}$ is

$$\frac{\lambda s}{\frac{s(s-2V)}{1 + \frac{\lambda C s}{s(s-2V)}}} = \frac{\lambda s}{s^2 + s(\lambda C - 2V)} = \frac{\lambda}{s + \lambda C - 2V} \quad (8.14)$$

It is then clear that since it is desired that the required thrust be a diminishing quantity, the system should be stable meaning that

$$\lambda C - 2V = \lambda \frac{V_{IR}}{R} + 2 \frac{\dot{R}}{R} > 0 \quad (8.15)$$

$$\text{or } \lambda > \frac{-2 \dot{R}/R}{V_{IR}/R} = - \frac{2 \dot{R}}{V_{IR}} \quad (8.16)$$

The quantity $\lambda = \frac{1}{-R/V_{IR}}$ is referred to as effective navigation constant and the condition for stability is then $\lambda > 2$. It is clearly seen that when the interceptor is on collision course and $a_{Tn} = 0$, the quantity $\dot{\lambda}$ remains zero for the entire flight which indicates that no acceleration is required.

9. Tracking Systems

In the preceding part, it was pointed out that for closed-loop guidance of vehicles, a measurement of either line-of-sight angle or angle rate with respect to a fixed reference is necessary. This job is accomplished by a tracking system (usually called angle tracking system).

A tracking system is composed of two parts: a) an error detector, which is a device producing a signal proportion to the deviation of antenna center beam from LOS and b) a control loop which drives the antenna in order to reduce the tracking error to zero. If the tracking error is maintained very small during the flight, the position of the antenna, which is known (can be measured) gives the information about δ or $\dot{\delta}$ (LOS angle or rate).

a) Error Detectors

The following are three important examples of error detectors.

Radar Lobing - Fig. 9.1 represents the propagation pattern of a radar antenna.

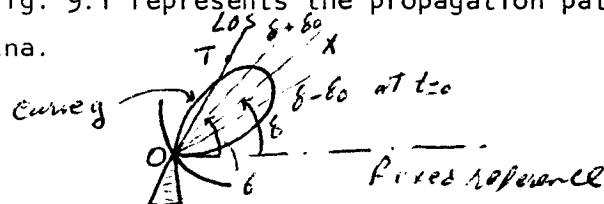


Fig. 9.1 - Radar Antenna

Since δ is known, the object is to generate a signal proportion to $\delta - \delta_0$ where T is the target point. When T is on ox a maximum

signal is returned. The farther T is from ox in angle, the smaller will be the amplitude of the return signal. This functional relationship is shown by the curve g in Fig. 9.1 and can be represented as

$$\beta = K g(\delta - \epsilon) \quad (9.1)$$

where z is the radar output and is the analytical representation of the lobe and K is proportion to the length OT . Now let us move the antenna mechanically, or electronically, by the following rule

$$\delta = \delta_0 + \delta_1 \sin \omega t \quad (9.2)$$

where δ_1 is a constant known quantity. It can be seen that

$$\beta = -K_1 (\delta - \epsilon_0) \sin \omega t \quad (9.3)$$

Multiplying z by $\sin \omega t$ and keeping the d.c. term yields

$$\beta_{d.c.} = -\frac{1}{2} K_1 (\delta - \epsilon_0) \quad (9.4)$$

This is the required result. Notice that $z_{d.c.}$ contains the sign of $\delta - \epsilon_0$, i.e. the position of T with respect to center of the lobe (δ_0).

Error Detection with Fixed Antennas - In certain applications, it is necessary not to mechanically, or electronically, lobe the antenna. In this case two antennas can be used to determine the phase difference between the received signals (Fig. 9.2).

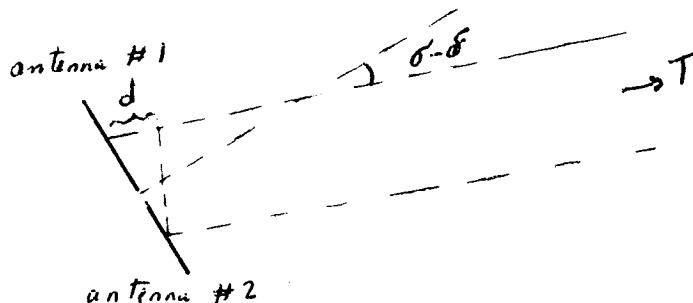


Fig. 9.2 - Error Detector without Lobing

It is easily seen that the distance d is proportion to $\epsilon' - \epsilon$ angle. The incident lines from target are almost parallel lines because the size of combination antennas is much smaller than the distance to target. The two outputs of the antennas are

$$z_1 = K \sin \omega t \quad (9.5)$$

$$z_2 = K \sin(\omega t + cd) \quad (9.6)$$

Phase shifting z_1 by 90° and then multiplying z_1 and z_2 and retaining the d.c. component yields

$$\text{d.c. component} = \frac{K^2}{2} \sin cd \quad (9.7)$$

For small angles ($\epsilon' - \epsilon$) approximately

d.c. comp. = proportional to $\epsilon' - \epsilon$. Again notice that the d.c. component in (9.7) retains the sign information in $\epsilon' - \epsilon$.

Optical Error Detector - The idea of lobing can be used in an optical detector. Let four detectors be placed as shown in Fig. 9.3.

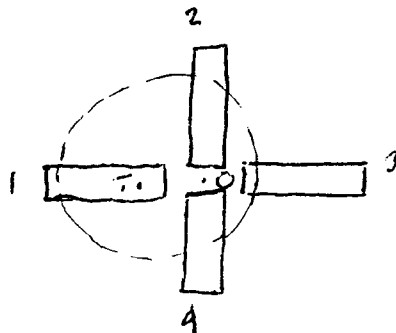


Fig. 9.3 - Optical Error Detector

Let the target image be at point T without lobing. Lobing therefore will move the image on a circle with center T . When the image is on a detector, an output is produced. The series combination of four outputs of the detectors for the case of Fig. 9.3 is presented by Fig. 9.4a. If T coincides with O the center of the detectors, then

the case of Fig. 9.4b will result. Consequently, information about angle of incident light from the target can be obtained from z

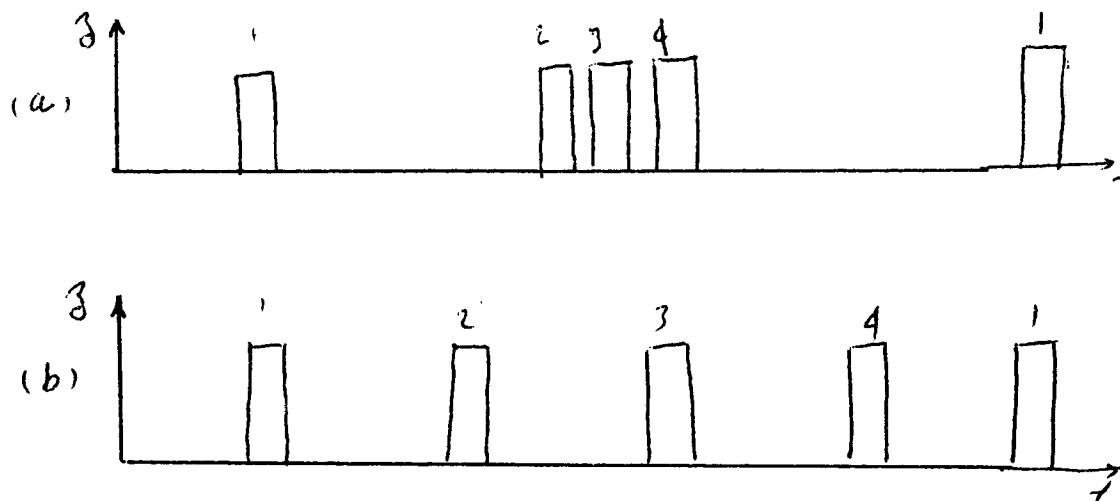


Fig. 9.4 - Output Waveform of Optical Detector

b) Tracking Loop.

Any of the above schemes will yield a relationship between the detector output and $\delta - \delta$ which typically can be represented as in Fig. 9.5

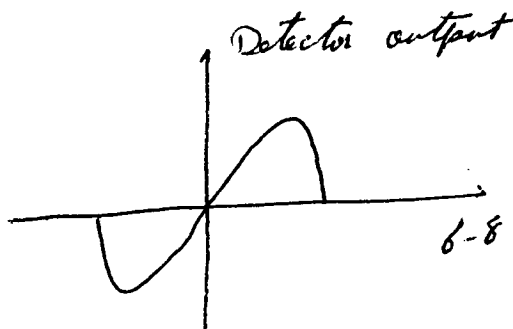


Fig. 9.5 - Characteristic of Error Detectors

The S-shaped form is because when $\delta - \delta$ is larger than certain value, the target falls completely outside the radar or optical beam.

Fig. 9.6 represents the block diagram of a complete tracking system

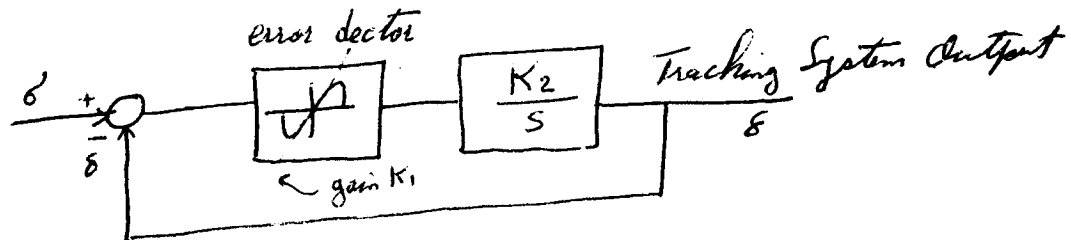


Figure 9.6 - Tracking System

over the linear range of the error detector, we have

$$\mathcal{L} \delta = \mathcal{L} \delta \frac{K_1 K_2 / s}{1 + K_1 K_2 / s} = \frac{K_1 K_2}{s + K_1 K_2} \quad (9.8)$$

For example if the antenna is initially not aligned with the target (i.e. $\delta \neq \delta'$) then from 9.8

$$\delta(t) = \delta' [1 - e^{-K_1 K_2 t}] \quad (9.9)$$

which shows that as $t \rightarrow \infty$, $\delta(t) \rightarrow \delta'$, the desired position. In practice, the terms K_2/s should be modified to include the antenna dynamics, and also any filtering of the noise, which may help the performance of the system.

243
163

N 6 7. 8. 0. 4 6. 6

SPACE SCIENCE -SPACE COMMUNICATIONS

V

by

R. Scholtz
C. Weber

Introduction

Communication system engineering is a demanding discipline. Theoretical design requires an extensive knowledge of almost every field of applied mathematics; practical design requires a knowledge of the operational characteristics of devices, the effects occurring in energy radiation, etc. Underlying all of these prerequisites to competence is a knowledge of statistics.

Statistics and its companion subject, probability, concern the measurement and description of non-deterministic phenomena. This subject must be broached since information (i.e. that quantity which is conveyed by the communication system) must be in essence non-deterministic. Everyone can picture the foolishness of designing a communication system to transmit the message "the sun will rise tomorrow." The user of the systems output would not gain any information when this astounding news reaches his ears. Thus, the user cannot know the system input in advance.

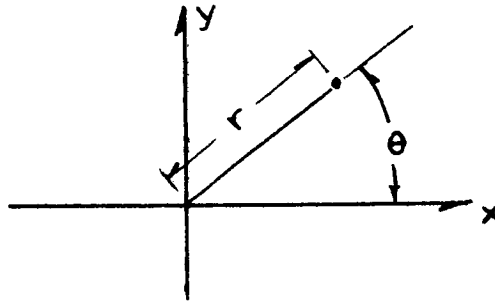
Interference in communication systems also demands statistical techniques for analysis. For example, other random signals in addition to the transmitted signal may enter the receiving antenna in a communication system. Since the receiver does not know the signal which the transmitter broadcasts, it may erroneously "interpret" the signal sensed by its antenna. Again an analysis of the effects of non-deterministic waveforms on receivers is necessary to evaluate receiver design.

It should be apparent now that the prerequisites must be studied before we can discuss system design problems intelligently. Part I of this course will provide a most elementary review of the mathematical tools required of the communication engineer. Part II will discuss the components of communication systems and the analysis of system performance.

PART I
MATHEMATICAL BACKGROUND

A. Complex Variables

A complex number Z is simply an ordered pair (x,y) of real numbers. As such Z is simply a point in the (x,y) plane.



Considering complex numbers as vectors, the sum of two complex numbers $Z_1 = (x_1, y_1)$ and $Z_2 = (x_2, y_2)$ given by addition of vector components

$$Z_1 + Z_2 = (x_1 + x_2, y_1 + y_2) \quad (1)$$

However, multiplication, which is not defined for vectors, is defined for complex numbers

$$Z_1 Z_2 = (x_1 x_2 - y_1 y_2, x_1 y_2 + x_2 y_1) \quad (2)$$

The additive identity or "zero" which leaves a complex number Z unchanged under addition is denoted by $0 = (0,0)$.

$$Z + 0 = (x + 0, y + 0) = Z \quad (3)$$

The multiplicative identity or "one" which leaves a complex number unchanged under multiplication is denoted by $1 = (1,0)$.

$$1 \cdot Z = (1 \cdot x - 0 \cdot y, 1 \cdot y + 0 \cdot x) = Z \quad (4)$$

We define $-Z$ as the number satisfying the equation

$$Z + (-Z) = 0 \quad (5)$$

Thus the additive inverse of Z is given by

$$-Z = (-x, -y) \quad (6)$$

Likewise Z^{-1} is the number satisfying the equation

$$ZZ^{-1} = 1 \quad (7)$$

It can be verified that

$$Z^{-1} = \left(\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2} \right) \quad (8)$$

The complex numbers can be derived from the solution to the equation

$$Z^2 = -1 \quad (9)$$

We all realize that no real number can satisfy this equation. However, two specific complex numbers can. Using the relations

$$Z^2 = (x^2 - y^2, 2xy) \quad , \quad -1 = (-1, 0) \quad (10)$$

and equating coefficients of these vectors, we see that the solutions to (10) are given by

$$Z = (0, 1) \quad \text{or} \quad (0, -1) \quad (11)$$

We shall denote $(0, 1)$ by i .

We see now that every complex number can be written as a linear combination of the multiplicative identity 1 and the square root i of -1 , where the coefficients are real numbers. Thus,

$$\begin{aligned}
 Z &= x(1,0) + y(0,1) \\
 &= x \cdot 1 + y \cdot i \\
 &= x + yi
 \end{aligned}
 \tag{12}$$

We often refer to x as the real part of Z and y as the imaginary part of Z .

As we shall soon see, the usefulness of complex numbers in electronic engineering depends on their multiplicative properties. The evaluation of products of complex numbers is performed most easily in polar coordinates. If

$$\begin{aligned}
 x &= r \cos \theta, & r &= \sqrt{x^2 + y^2} \\
 y &= r \sin \theta, & \tan \theta &= y/x
 \end{aligned}
 \tag{13}$$

Then

$$Z = x + yi = r(\cos \theta + i \sin \theta)
 \tag{14}$$

We call r the magnitude or absolute value of Z and write

$$r = |Z|$$

Likewise, we term θ the phase or angle of Z .

Using the series expansions for $\sin \theta$ and $\cos \theta$, we see that equation (14) reduces to

$$\begin{aligned}
 Z &= r \left[\left(\sum_{n=0}^{\infty} (-1)^n \frac{\theta^{2n}}{(2n)!} \right) + i \left(\sum_{n=0}^{\infty} (-1)^n \frac{\theta^{2n+1}}{(2n+1)!} \right) \right] \\
 &= r \left[\sum_{m=0}^{\infty} \frac{(i\theta)^m}{m!} \right] = r e^{i\theta}
 \end{aligned}
 \tag{15}$$

The rules of multiplication simplify considerably since for

$$\begin{aligned}
 Z_j = (x_j, y_j) &= r_j e^{i\theta_j} & \text{equation (10) reduces to} \\
 Z_1 Z_2 &= r_1 r_2 e^{i(\theta_1 + \theta_2)}
 \end{aligned}
 \tag{16}$$

Multiplication corresponds to multiplying magnitudes and adding phase angles. The inverse of Z reduces to

$$Z^{-1} = \frac{1}{r} e^{-i\theta} \quad (17)$$

We can define the conjugate of Z as $Z^* = r e^{-i\theta}$. Then

$$Z Z^* = (r e^{i\theta})(r e^{-i\theta}) = r^2 = |Z|^2 \quad (18)$$

Obviously conjugation can be regarded as changing i to $-i$ in an expression. Thus

$$(Z_1 Z_2)^* = Z_1^* Z_2^* \quad (19)$$

B. Fourier Series

Definition: We say that a function $g(t)$ is periodic if there exists a constant T such that for all t ,

$$g(t) = g(t+T) \quad (20)$$

Consider now the complex function of time

$$e^{i\omega_0 t} = \cos \omega_0 t + i \sin \omega_0 t \quad (21)$$

where ω_0 is a constant. Certainly $e^{i\omega_0 t}$ is a periodic function since for any value of t ,

$$e^{i\omega_0 t} = e^{i\omega_0(t + \frac{2\pi}{\omega_0})} = e^{i\omega_0(t+T)} \quad (22)$$

where $T = \frac{2\pi}{\omega_0}$ is known as the period of the function.

Theorem: Let $g(t)$ be any real or complex periodic function of period T , satisfying the condition

$$\int_0^T |g(t)|^2 dt < \infty \quad (23)$$

Then $g(t)$ can be represented by the Fourier series expansion

$$\boxed{g(t) = \sum_{n=-\infty}^{\infty} a_n e^{in\omega_0 t}} \quad (24)$$

where $\omega_0 = \frac{2\pi}{T}$ and the a_n are complex constants.

This theorem simply admits that any periodic function of time can be represented by a sum of elementary periodic functions of time, namely $e^{in\omega_0 t}$. We call $e^{in\omega_0 t}$ the n^{th} harmonic of $x(t)$.

Suppose we were given a periodic function satisfying (22). The coefficients a_n in the Fourier series expansion (24) can be determined via the following integral:

$$\begin{aligned} \int_0^T e^{-in\omega_0 t} g(t) dt &= \int_0^T e^{-in\omega_0 t} \left[\sum_{m=-\infty}^{\infty} a_m e^{im\omega_0 t} \right] dt \\ &= \sum_{m=-\infty}^{\infty} a_m \left[\int_0^T e^{i(m-n)\omega_0 t} dt \right] \end{aligned} \quad (25)$$

If we expand the integral in (25) we have that

$$\int_0^T e^{i(m-n)\omega_0 t} dt = \int_0^T [\cos(m-n)\omega_0 t + i \sin(m-n)\omega_0 t] dt \quad (26)$$

Using the fact that the integral of a cosine or sine wave over an integer number of periods is zero, we see that

$$\int_0^T e^{i(m-n)\omega_0 t} dt = \begin{cases} 0, & m \neq n \\ T, & m = n \end{cases} = T \delta_{mn} \quad (27)$$

The function δ_{mn} which is 0 for $m \neq n$ and 1 for $m = n$ is known as the Kronecker delta function. By substituting (27) back into (25) we see that

$$\int_0^T g(t) e^{-in\omega_0 t} dt = \sum_{m=-\infty}^{\infty} a_m [T \delta_{mn}] = a_n T \quad (28)$$

and thus the Fourier series coefficients are given by

$$a_n = \frac{1}{T} \int_0^T g(t) e^{-in\omega_0 t} dt \quad (29)$$

for all values of n .

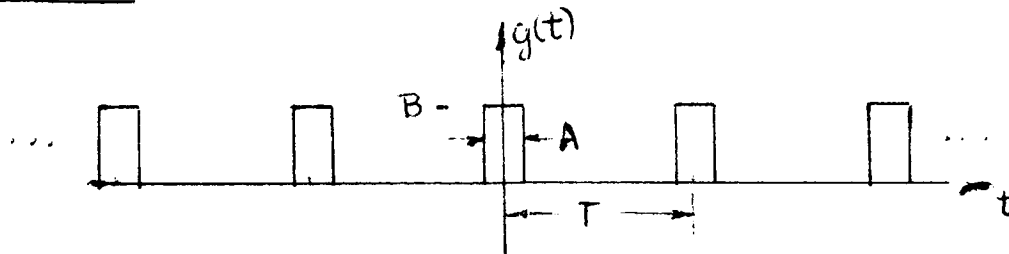
It is interesting to note that the actual value of the integral (23) can be determined very easily from the series coefficients a_n since, using (19), (24) and (27),

$$\begin{aligned} \int_0^T |g(t)|^2 dt &= \int_0^T \left[\sum_{n=-\infty}^{\infty} a_n e^{in\omega_0 t} \right] \left[\sum_{m=-\infty}^{\infty} a_m^* e^{-im\omega_0 t} \right] dt \\ &= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_n a_m^* \int_0^T e^{i(n-m)\omega_0 t} dt \\ &= T \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_n a_m^* \delta_{mn} \end{aligned} \quad (30)$$

Thus we have derived Parseval's Theorem for Fourier series which states that

$$\sum_{n=-\infty}^{\infty} |a_n|^2 = \frac{1}{T} \int_0^T |g(t)|^2 dt \quad (31)$$

Example 1: Consider the periodic function $g(t)$ as shown below:



Here A and B are real numbers. Then

$$\int_0^T |g(t)|^2 dt = B^2 A < \infty$$

Since (23) is satisfied, $g(t)$ has a Fourier series expansion with coefficients given by (29).

$$a_n = \frac{1}{T} \int_0^T g(t) e^{-in\omega_0 t} dt$$

equation (29).

$$= \frac{1}{T} \int_{-T/2}^{T/2} g(t) e^{-in\omega_0 t} dt$$

since integrand has period T

$$= \frac{B}{T} \int_{-A/2}^{A/2} e^{-in\omega_0 t} dt$$

definition of $g(t)$

$$= \frac{B}{T} \left. \frac{e^{-in\omega_0 t}}{-in\omega_0} \right|_{-A/2}^{A/2}$$

performing integration

$$= \frac{TB}{\pi n} \sin(\pi n A/T)$$

since $\omega_0 = 2\pi/T$
and $\sin \alpha = \frac{e^{i\alpha} - e^{-i\alpha}}{2i}$

Thus
$$g(t) = \frac{B}{\pi} \sum_{n=-\infty}^{\infty} \frac{\sin(\pi n A/T)}{n} e^{in\omega_0 t}$$

From our study of Fourier series, it is now obvious that there exists a one-to-one relationship between the periodic function $g(t)$ and the sequence a_n of Fourier coefficients. In many cases, our main interest is not in the value of the coefficients a_n , but merely in their squared magnitude $|a_n|^2$. If for example, $g(t)$ denotes a voltage or velocity, $|a_n|^2$ will be proportional to the power in $g(t)$ at frequency n/T .

Definition: The power spectral density $S_g(f)$ of a periodic function $g(t)$ is given by

$$S_g(f) = \sum_{n=-\infty}^{\infty} |a_n|^2 \delta(f - \frac{n}{T}) \quad (32)$$

where $\delta(f - \frac{n}{T})$ denotes a Dirac delta function at frequency n/T . The Dirac delta function mentioned here is simply a "very narrow" function with area 1, having essentially all of this area at the point where the argument of the function is zero. Thus for practical purposes

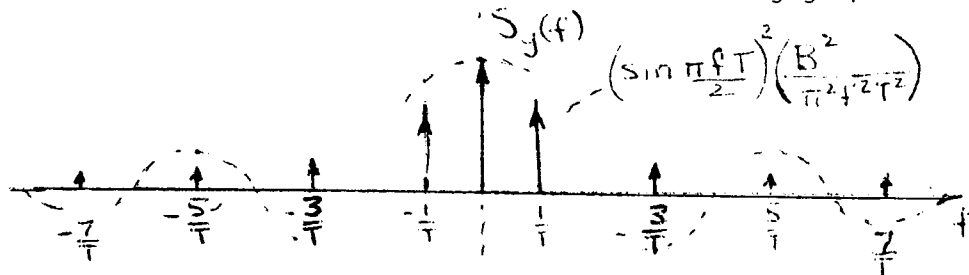
$$\delta(f - \alpha) = \begin{cases} 0 & \text{for } f \neq \alpha \\ \infty & \text{for } f = \alpha \end{cases} \quad (33)$$

$$\int_a^b S(f-\alpha) df = \begin{cases} 1 & \text{if } a < \alpha < b \\ 0 & \text{if } \alpha < a \text{ or } \alpha > b \end{cases} \quad (34)$$

Example 2: The spectral density of ^{the} square wave of example 1 is given by

$$S_g(f) = \sum_{n=-\infty}^{\infty} \left| \frac{B \sin(n\pi A/T)}{n\pi} \right|^2 \delta(f - \frac{n}{T})$$

For $A = T/2$, $B = 2$, this function has the following graph:



The spectral density $S_g(f)$ defined above is known as a two-sided spectral density, the reason being that $S_g(f)$ has values for negative frequency. This should not disturb you since we know that $\cos(-\omega t) = \cos(\omega t)$ and $\sin(-\omega t) = -\sin \omega t$. Thus $|a_n|^2 + |a_{-n}|^2$ indicates the total realizable power at frequency n/T . The use of two-sided spectral densities is simply a mathematical convenience.

To obtain the total power at any set of frequencies, we need only to integrate the power spectral density over the frequency range of interest. Since Dirac delta functions integrate to one, this integration simply adds their coefficients $|a_n|^2$. Thus the integral of the power spectral density over the complete frequency range should be equal to the total power in the process.

$$\begin{aligned} \int_{-\infty}^{\infty} S_g(f) df &= \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} |a_n|^2 \delta(f - n/T) df \\ &= \sum_{n=-\infty}^{\infty} |a_n|^2 \end{aligned} \quad (35)$$

By Parseval's theorem (31), this is equal to the total power in the process.

Another method of obtaining a function which contains information only about $|a_n|^2$, involves the evaluation of an integral.

Definition: The time correlation function $\mathcal{K}_g(\tau)$ of a function $g(t)$ is given by the integral

$$\mathcal{K}_g(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T g(t+\tau) g^*(t) dt \quad (36)$$

Suppose we now calculate the time correlation function of a periodic function. Substituting (24) into (36) gives

$$\begin{aligned} \mathcal{K}_g(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \left[\sum_{n=-\infty}^{\infty} a_n e^{in\omega_0(t+\tau)} \right] \left[\sum_{m=-\infty}^{\infty} a_m e^{im\omega_0 t} \right]^* dt \\ &= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_n a_m^* e^{in\omega_0 \tau} \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T e^{i(n-m)\omega_0 t} dt \end{aligned} \quad (37)$$

It can be easily verified with the aid of (27) that the above integral is equal to the δ_{mn} , the Kronecker delta function. Thus

$$\mathcal{K}_g(\tau) = \sum_{n=-\infty}^{\infty} |a_n|^2 e^{in\omega_0 \tau} \quad (38)$$

It is now obvious that if we expand the correlation function in a Fourier series in τ , the Fourier coefficients of this expansion will be $|a_n|^2$. While this approach to determining $|a_n|^2$ appears longer and more complicated than a direct evaluation of the coefficients, the result demonstrated in (38) can be extended to non-deterministic and non-periodic situations as we shall soon see.

C. Fourier Transforms

We can think of a non-periodic function as the limiting case of a periodic function whose period T has come arbitrarily large. If we examine the Fourier series representation as $T \rightarrow \infty$, we see that (24) and (29) can be interpreted as follows:

$$g(t) = \sum_{n=-\infty}^{\infty} T a_n e^{i 2\pi n t / T} \xrightarrow{T \rightarrow \infty} \int_{-\infty}^{\infty} G(f) e^{i 2\pi f t} df$$

$$T a_n = \int_{-T/2}^{T/2} g(t) e^{-i 2\pi n t / T} dt \xrightarrow{T \rightarrow \infty} G(f) = \int_{-\infty}^{\infty} g(t) e^{-i 2\pi f t} dt \quad (39)$$

Here we have used the fact that $\omega_n = \frac{2\pi}{T}$, and defined the limiting values as $n/T \rightarrow f$, $1/T \rightarrow df$, and $T a_n \rightarrow G(f)$ in converting the Fourier series to an integral. The pair of equations (39) are simply plausibility arguments for the following theorem which we shall now state.

Theorem: If a function $g(t)$ satisfies the absolute integrability relation

$$\int_{-\infty}^{\infty} |g(t)| dt < \infty \quad (40)$$

then there exists a unique function $G(f)$ such that $g(t)$ and $G(f)$ satisfy the following relations:

$$\boxed{\begin{aligned} G(f) &= \int_{-\infty}^{\infty} g(t) e^{-i \omega t} dt \\ g(t) &= \int_{-\infty}^{\infty} G(f) e^{i \omega t} df \end{aligned}} \quad (41)$$

where $\omega = 2\pi f$.

A list of Fourier transforms is given in Table 1.

Table 1

$g(t)$	$G(f)$	
$\delta(t)$	1	(42)
1	$\delta(f)$	(43)
$e^{i\omega_0 t}$	$\delta(f-f_0)$	(44)
$\delta(t-t_0)$	$e^{-i\omega t_0}$	(45)
$\cos \omega_0 t$	$\frac{1}{2} [\delta(f-f_0) + \delta(f+f_0)]$	(46)
$\sin \omega_0 t$	$\frac{1}{2i} [\delta(f-f_0) - \delta(f+f_0)]$	(47)
$\frac{1}{T_0} U(t) e^{-t/T_0}$	$\frac{1}{1+i\omega T_0}$	(48)
$e^{-a t }$	$\frac{2a}{a^2 + \omega^2}$	(49)
$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{t-m}{\sigma}\right)^2\right\}$	$\exp\{-i\omega m - \omega^2 \sigma^2/2\}$	(50)
$\frac{1}{2T} 2T (t)$	$\frac{\sin 2\pi T f}{2\pi T f}$	(51)
$\frac{\sin 2\pi W t}{2\pi W t}$	$\frac{1}{2W} 2W (f)$	(52)

In this table we define a useful notation for two functions:

$$U(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (53)$$

$$|T|(t) = \begin{cases} 1 & -\frac{T}{2} \leq t \leq \frac{T}{2} \\ 0 & \text{elsewhere} \end{cases} \quad (54)$$

The usefulness of the transform concept relies on the fact that manipulations which are difficult in the time domain are simple in the frequency domain. This statement will become apparent in the course of our studies. There is a variety of theorems which indicate the effects of time domain manipulations on transforms. For example, let $g_1(t)$ and $g_2(t)$ be absolutely integrable functions whose transforms are $G_1(f)$ and $G_2(f)$ respectively. Then consider the function

$$h(t) = b_1 g_1(t) + b_2 g_2(t) \quad (55)$$

Notice that

$$\int_{-\infty}^{\infty} |h(t)| dt \leq |b_1| \int_{-\infty}^{\infty} |g_1(t)| dt + |b_2| \int_{-\infty}^{\infty} |g_2(t)| dt < \infty \quad (56)$$

since $g_1(t)$ and $g_2(t)$ are absolutely integrable. Therefore $h(t)$ has a transform given by (41).

$$\begin{aligned} H(f) &= \int_{-\infty}^{\infty} h(t) e^{-i\omega t} dt \\ &= b_1 \int_{-\infty}^{\infty} g_1(t) e^{-i\omega t} dt + b_2 \int_{-\infty}^{\infty} g_2(t) e^{-i\omega t} dt \\ &= b_1 G_1(f) + b_2 G_2(f) \end{aligned} \quad (57)$$

Equation (57) describes the effect in the frequency domain of the manipulation (55) in the time domain.

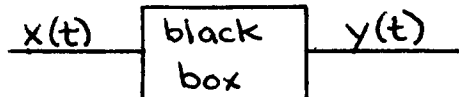
Applying the transform relation (55) - (57) to the correlation function of a periodic function given by (38) in the previous chapter, we see that the transform of the time correlation function of $g(t)$ is the spectral density of $g(t)$. Thus

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathcal{R}_g(\tau) e^{-i\omega\tau} d\tau &= \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} |a_n|^2 e^{-i(\omega - n\omega_0)\tau} d\tau \\
&= \sum_{n=-\infty}^{\infty} |a_n|^2 \int_{-\infty}^{\infty} e^{-i(\omega - n\omega_0)\tau} d\tau \\
&= \sum_{n=-\infty}^{\infty} |a_n|^2 \delta(f - nf_0) = S_g(f) \quad (58)
\end{aligned}$$

Thus the Fourier transform is useful in deriving spectral densities from time correlation functions. We shall introduce more transform manipulation theorems as they are required.

D. Linear Operations

Engineers are notorious for thinking in terms of system block diagrams. Suppose we examine the simplest "black box" diagram to determine if there is a simple way to describe the system.



The system operates on the input $x(t)$ to give the output $y(t)$. If the operation performed by the system has the following two properties, its description can be significantly simplified.

Definition: A system is said to be linear if for any two inputs $x_1(t)$ and $x_2(t)$ which cause outputs $y_1(t)$ and $y_2(t)$, the composite signal $a_1x_1(t) + a_2x_2(t)$ causes output $a_1y_1(t) + a_2y_2(t)$, a_1 and a_2 being arbitrary constants.

Definition: A system is said to be time-invariant if for any input $x(t)$ which causes output $y(t)$, the input $x(t + \tau)$ causes output $y(t + \tau)$ for all τ .

As far as these definitions are concerned, $x(t)$, $y(t)$ and the constants can all be complex functions, though in most cases the inputs and outputs are real functions of time.

Suppose that the input $x(t)$ is given by

$$x(t) = e^{i\omega_0 t} \quad (59)$$

and the output is $y(t)$. If we apply the constraint that the system be linear, then when the input is given by

$$x_1(t) = (e^{i\omega_0 \tau}) e^{i\omega_0 t} \quad (60)$$

the output is given by

$$y_1(t) = e^{i\omega_0 \tau} y(t) \quad (61)$$

On the other hand, if the system is time invariant, then for input

$$x_2(t) = e^{i\omega_0(t+\tau)} \quad (62)$$

the output is

$$y_2(t) = y(t+\tau) \quad (63)$$

Since the inputs (60) and (62) are identical in the two situations considered, the outputs (61) and (63) must also be identical. Thus we have

$$y(t+\tau) = e^{i\omega_0 \tau} y(t) \quad (64)$$

The relation (64) is the basis of "steady state" linear circuit analysis. To clarify this statement, let us define

$$y(o) = H(f_o) \quad (65)$$

where $f_o = \omega_o/2\pi$. (Certainly the response $y(t)$ depends on the input frequency chosen in (59).) If we then evaluate (64) at $t = 0$, and then perform a change of variables $\tau = t$, the result is

$$y(t) = e^{i\omega_0 t} H(f_o) = H(f_o) x(t) \quad (66)$$

Thus $y(t)$ is obtained from the input exponential harmonic function $e^{i\omega_0 t}$ by a simple multiplication of $x(t)$ by the complex number $H(f_0)$. The magnitude of $H(f_0)$ is called the gain of the system at frequency f_0 , and the argument (or angle) of $H(f_0)$ is called the phase-shift of the system at frequency f_0 .

Suppose we now examine the operation of a time-invariant linear system when the input is a Fourier transformable function $x(t)$ with transform $X(f)$. We can now write an expression for $x(t)$ involving exponential functions

$$x(t) = \int_{-\infty}^{\infty} e^{i\omega t} X(f) df = \lim_{\Delta f \rightarrow 0} \left\{ \Delta f \sum_{n=-\infty}^{\infty} e^{in2\pi\Delta f t} X(n\Delta f) \right\} \quad (67)$$

We have now written $x(t)$ as the limiting case of a linear combination of exponential functions. Assuming that the system function is denoted by $H(f)$ and using the linear properties of the system, we know that the input $e^{in2\pi\Delta f t} X(n\Delta f)$ gives output $e^{in2\pi\Delta f t} X(n\Delta f) H(n\Delta f)$ and thus the total output $y(t)$ is given by

$$\begin{aligned} y(t) &= \lim_{\Delta f \rightarrow 0} \left\{ \sum_{n=-\infty}^{\infty} e^{in2\pi\Delta f t} X(n\Delta f) H(n\Delta f) \Delta f \right\} \\ &= \int_{-\infty}^{\infty} X(f) H(f) e^{i\omega t} df \end{aligned} \quad (68)$$

or equivalently the transform of the output is given by

$$\boxed{Y(f) = X(f) H(f)} \quad (69)$$

The transform of the output of a linear system is simply the product of the transform of the input and the system function.

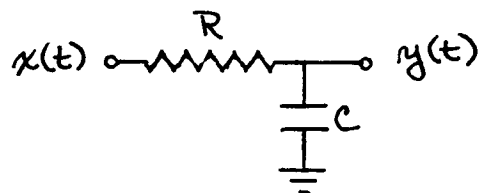
Equation (69) implies a simple relation in the time domain. If we retreat to (68) by transforming both sides of (69) and then substitute a transform integral for $X(f)$, we have

$$\begin{aligned}
 y(t) &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x(t') e^{-i\omega t'} dt' \right] H(f) e^{i\omega t} df \\
 &= \int_{-\infty}^{\infty} x(t') \left[\int_{-\infty}^{\infty} H(f) e^{i\omega(t-t')} df \right] dt' \\
 &= \int_{-\infty}^{\infty} x(t') h(t-t') dt'
 \end{aligned}$$

(70)

where $h(t)$ is the inverse Fourier transform of the system function $H(f)$. The integral in (70) is known as the convolution integral and arises whenever transforms are multiplied as in (69). The time function $h(t)$ is known as the impulse response of the system for the following reason. If the input $x(t)$ is a delta function (or impulse) $\delta(t)$, then $X(f) = 1$ and $Y(f) = H(f)$. Taking inverse transforms of both sides gives $y(t) = h(t)$ and thus $h(t)$ is the output of the system when the input is an impulse.

Example 3: Suppose we consider an R-C filter



with $x(t)$ representing the applied voltage and $y(t)$ the output voltage under no load conditions. The operation of this circuit is governed by the differential equation for the input current.

$$[x(t) - y(t)] \frac{1}{R} = C \frac{dy}{dt} \quad (71)$$

We can now transform both sides of the equation (71) if we can determine the transform of the derivative of $y(t)$. Using integration by parts:

$$\int_{-\infty}^{\infty} \left(\frac{dy}{dt} \right) e^{-i\omega t} dt = y(t) e^{-i\omega t} \Big|_{-\infty}^{\infty} + (i\omega) \int_{-\infty}^{\infty} y(t) e^{-i\omega t} dt \quad (72)$$

Using the fact that if $x(t) \rightarrow 0$ as $t \rightarrow \pm \infty$ since $x(t)$ is assumed absolutely integrable, then $y(\pm \infty) = 0$, we see that

$$\boxed{\int_{-\infty}^{\infty} \left(\frac{dy}{dt} \right) e^{-i\omega t} dt = (i\omega) Y(f)} \quad (73)$$

since the integral in (72) is simply the transform of $y(t)$. Thus the transform of (71) is given by

$$Y(f) = \frac{1}{1 + (RC)(i\omega)} X(f) \quad (74)$$

Since (71) is a linear differential equation with constant coefficients, (74) is equivalent to (69) and the system function of the R-C filter is

$$H(f) = \frac{1}{1 + (RC)(i\omega)} \quad (75)$$

From (48) in the table of Fourier transform relations, the impulse response of the filter is

$$h(t) = \frac{1}{RC} U(t) e^{-t/RC} \quad (76)$$

and generally the output of the R-C filter can be written as

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} x(t') \frac{1}{RC} U(t-t') e^{-(t-t')/RC} dt' \\ &= \frac{e^{-t/RC}}{RC} \int_{-\infty}^t x(t') e^{t'/RC} dt' \end{aligned} \quad (77)$$

using the convolution theorem (69)-(70).

E. A Sampling Theorem

One particular process which occurs in practice is that of sampling. We shall regard sampling as a delta function modulation process. Thus the sampled values of the wave $g(t)$ are denoted by

$$g_s(t) = \sum_{n=-\infty}^{\infty} g(nT) \delta(t-nT) = g(t) \left[\sum_{n=-\infty}^{\infty} \delta(t-nT) \right] \quad (78)$$

In this notation we mean that the area of the delta function $g(nT)\delta(t-nT)$ occurring at time nT is $g(nT)$.

Suppose we assume that $g_s(t)$ is Fourier transformable and compute its transform. To do this we shall apply the analog to the convolution theorem ((69)-(70)), which can be derived by simply interchanging f and t , and i and $-i$ in (70). The theorem states that if we have the transform pairs

$$x(t) \rightarrow X(f)$$

$$y(t) \rightarrow Y(f)$$

then

$$\mathcal{F}\{x(t)y(t)\} = \int_{-\infty}^{\infty} X(f-\alpha)Y(\alpha)d\alpha \quad (79)$$

Thus, multiplication in the time domain corresponds to convolution in the frequency domain.

To apply (79) to the right hand side of (78), we must determine the Fourier transform of a delta function train. A simple form for the answer is obtained if we first expand the δ function train in a Fourier series since it is periodic. By (24) and (29) then, we have

$$\sum_{n=-\infty}^{\infty} [\delta(t-nT)] = \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{in2\pi t/T} \quad (80)$$

Taking Fourier transforms of both sides gives (using (44)):

$$\mathcal{F}\left\{\sum_{n=-\infty}^{\infty} [\delta(t-nT)]\right\} = \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta(f - \frac{n}{T}) \quad (81)$$

Equation (81) gives the amazing result that a δ function train transforms into another delta function train.

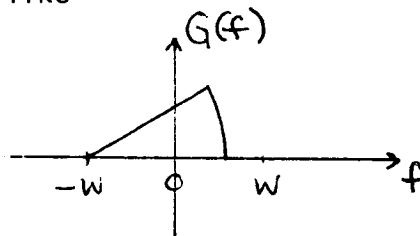
We can now use (79) and (81) to transform (78):

$$G_s(f) = \int_{-\infty}^{\infty} G(f-\alpha) \left[\frac{1}{T} \sum_{n=-\infty}^{\infty} \delta(\alpha - \frac{n}{T}) \right] d\alpha \quad (82)$$

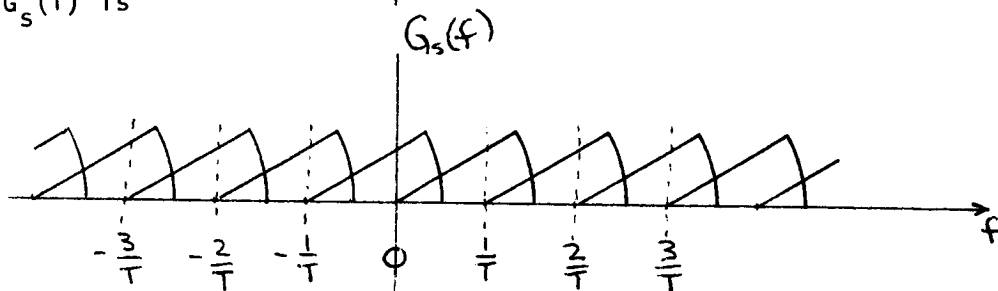
where $G_s(f)$ and $G(f)$ denote the transforms of $g_s(t)$ and $g(t)$ respectively. Performing the integration gives:

$$\begin{aligned}
 G_s(f) &= \frac{1}{T} \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} G(f-\alpha) \delta(\alpha - \frac{n}{T}) d\alpha \\
 &= \frac{1}{T} \sum_{n=-\infty}^{\infty} G(f - \frac{n}{T}) \int_{-\infty}^{\infty} \delta(\alpha - \frac{n}{T}) d\alpha \\
 &= \frac{1}{T} \sum_{n=-\infty}^{\infty} G(f - \frac{n}{T})
 \end{aligned} \tag{83}$$

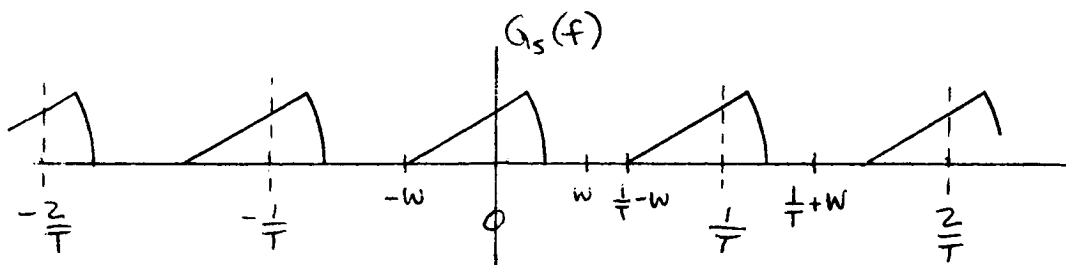
Thus if $G(f)$ looks like



then $G_s(f)$ is



An interesting effect occurs when $G_s(f)$ is bandlimited, i.e. $G(f) = 0$ for $|f| > W$. Then if $1/T > 2W$, the terms in (83) do not overlap and $G_s(f)$ looks like



Notice that in the frequency range $|f| < W$, $G_s(f) = G(f)$, and outside that range $G(f) = 0$. This implies that by use of a zonal filter we can recover $g(t)$ exactly since

$$G(f) = T \left[\frac{1}{T} \right](f) G_s(f) \quad \text{if } T < \frac{1}{2W} \quad (84)$$

Using the convolution theorem and (52) gives

$$\begin{aligned} g(t) &= \int_{-\infty}^{\infty} \frac{\sin \pi \frac{(t-\alpha)}{T}}{\frac{\pi(t-\alpha)}{T}} \left[\sum_{n=-\infty}^{\infty} g(nT) \delta(\alpha - nT) \right] d\alpha \\ &= \sum_{n=-\infty}^{\infty} g(nT) \frac{\sin \left[\frac{\pi}{T}(t - nT) \right]}{\frac{\pi}{T}(t - nT)}, \quad \text{for } T < \frac{1}{2W} \end{aligned} \quad (85)$$

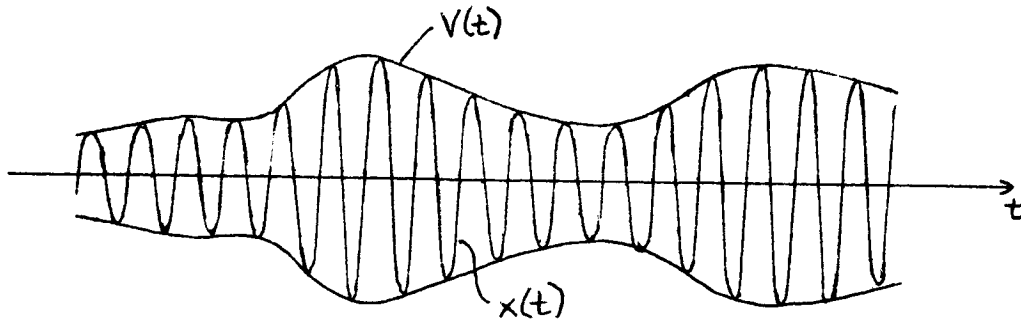
It is worth noting that to estimate $g(t)$ at an arbitrary value of t , one must have all the samples of the process. This is due to the fact that the filter $T \left[\frac{1}{T} \right](f)$ is not casual, and cannot be implemented in any physical system. Equation (85) can be approximated by physical systems, and in addition, is a useful representation for signals in analytic studies.

F. Non-Linear Operations

A real function $x(t)$ can always be expressed in the form

$$x(t) = V(t) \cos[\omega_c t + \phi(t)] \quad (86)$$

where $V(t) \geq 0$. In the case where both $V(t)$ and $\phi(t)$ are "slowly varying" with respect to a cosine wave of frequency ω_c , the quantities $V(t)$ and $\phi(t)$ have a physical interpretation (see figure). The quantity



$V(t)$ can be considered to be the envelope of $x(t)$, which can be constructed by connecting the peaks of $x(t)$ with a smooth curve. Likewise the quantity $\phi(t)$ can be considered to be the phase difference between a cosine wave of frequency f_0 cps. and the cosine wave in $x(t)$. We could equivalently write

$$x(t) = \text{Re} \{ V(t) e^{i[\omega_c t + \phi(t)]} \} \quad (87)$$

where $\text{Re} \{ \cdot \}$ denotes the real part of $\{ \cdot \}$. Functions where $V(t)$ and $\phi(t)$ have physical interpretations as envelope and phase, are called narrowband functions. Most transmitted signals in communication systems are narrowband.

Using the envelope-phase representation, it is possible to analyze the output of a class of zero-memory non-linear devices.

Definition: A zero-memory device has the property that its output at time t depends only on its input at time t .

Thus for a zero-memory device with input $x(t)$ and output $y(t)$ we have

$$y(t) = g(x(t)) \quad (88)$$

where $g(\cdot)$ is a known function. As usual the concept of a transform is quite useful here. Since the description of the device does not depend on time, we can state

$$y(t) = g(x(t)) = \int_{-\infty}^{\infty} G(f) e^{i\omega x(t)} df \quad (89)$$

where the transform relations is between the x and f domains. Let us make the substitution

$$y(t) = \int_{-i\infty}^{i\infty} G\left(\frac{s}{2\pi i}\right) e^{s x(t)} \frac{ds}{2\pi i} \quad (90)$$

The original constraint on this technique was that the function $g(\cdot)$ is Fourier transformable. In (90) we have converted the problem to two-sided Laplace transform notation. We can expand the exponential using the Jacobi-Anger formula

$$\exp\{z \cos \theta\} = \sum_{m=0}^{\infty} \epsilon_m I_m(z) \cos m\theta \quad (91)$$

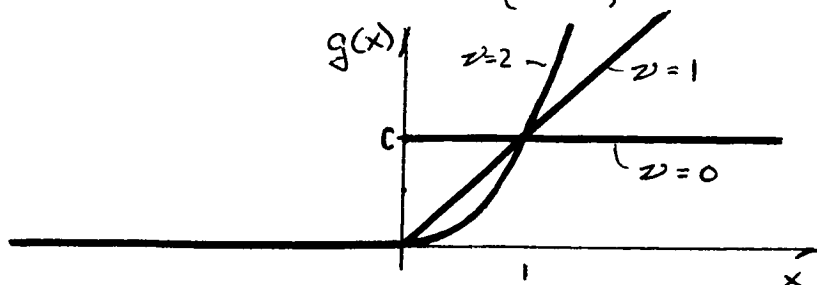
where $I_m(z)$ is the modified Bessel function and $\epsilon_m = \begin{cases} 2, & m \neq 0 \\ 1, & m = 0 \end{cases}$. Of course in applying (91) to expand e^{sx} we shall use the envelope-phase representation for x . Thus (90) reduces to

$$y(t) = \sum_{m=0}^{\infty} \epsilon_m \cos [m(\omega_c t + \phi(t))] \int_{-i\infty}^{i\infty} I_m(sV(t)) G\left(\frac{s}{2\pi i}\right) \frac{ds}{2\pi i} \quad (92)$$

We cannot proceed further without defining the zero memory device characteristic $g(\cdot)$. It is worth noting that we have in some sense created a harmonic expansion of $Y(t)$ with the m^{th} term in (90) denoting a signal with power in a frequency range around $\frac{m\omega_c}{2\pi}$ cps.

An important class of zero-memory devices are the half-wave z^{th} law devices which have the following characteristics:

$$y = g(x) = \begin{cases} 0, & x < 0 \\ cx^z, & x \geq 0 \end{cases} \quad (93)$$



The z^{th} law device characteristic is not directly Fourier transformable but by shifting the line of integration slightly to the right of the imaginary axis (see (92)), a legitimate representation for the transform $G(\frac{s}{2\pi i})$ is given by

$$G(\frac{s}{2\pi i}) = \frac{c \Gamma(z+1)}{s^{z+1}} \quad (94)$$

where $\Gamma(z+1)$ is the gamma function. If we make this substitution in (92), and let $s' = sV$, we have

$$y(t) = \sum_{m=0}^{\infty} C(z, m) V^z(t) \cos [m(\omega_0 t + \phi(t))] \quad (95)$$

where

$$C(z, m) = \frac{c \epsilon_m \Gamma(z+1)}{2\pi i} \int_{\epsilon-i\infty}^{\epsilon+i\infty} \frac{I_m(s')}{s'^{z+1}} ds' \quad (96)$$

for any $\epsilon > 0$.

The constant $C(z, m)$ can be evaluated as:

$$C(z, m) = \frac{c \epsilon_m \Gamma(z+1)}{2^{z+1} \Gamma(1 - \frac{m-z}{2}) \Gamma(1 + \frac{m+z}{2})} \quad (97)$$

The gamma function has the properties:

$$\begin{aligned} \Gamma(x) &= \int_0^{\infty} e^{-t} t^{x-1} dt \\ \Gamma(x+1) &= x \Gamma(x) \\ \Gamma(1) &= 1 \\ \Gamma(\frac{1}{2}) &= \sqrt{\pi} \end{aligned} \quad \left. \begin{aligned} \Gamma(n+1) &= n! \\ \Gamma(n+\frac{1}{2}) &= \frac{\sqrt{\pi} (2n)!}{2^{2n} n!} \end{aligned} \right\} \begin{array}{l} \text{for } n \\ \text{a positive} \\ \text{integer} \end{array}$$

$$\Gamma(n) = \infty \text{ for } n=0, -1, -2, -3, \dots$$

(98)

Using the last property in (98), we see that for $\frac{m-z}{2}$ a positive integer, $\Gamma(1 - \frac{m-z}{2})$ is infinite and $\zeta(z, m) = 0$.

Many practical devices do not have zero outputs when $x(t) < 0$.

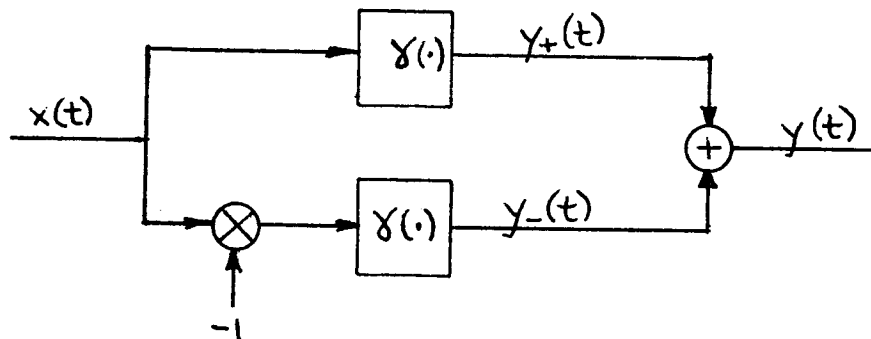
Definition: A full-wave even device is one whose characteristic satisfies the relation

$$g(x) = g(-x)$$

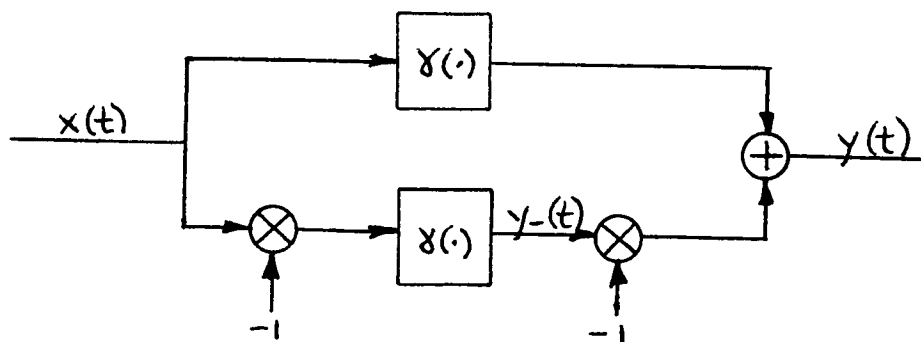
A full-wave odd device is one which satisfies

$$g(x) = -g(-x)$$

Suppose we let $\gamma(x) = U(x)g(x)$ be identical to $g(x)$ for $x > 0$ and zero otherwise. We can indicate full-wave even and odd devices by the following diagram:



full-wave even device, $g(x) = \gamma(x) + \gamma(-x)$



full-wave odd device, $g(x) = \gamma(x) - \gamma(-x)$

For ν^{th} law devices, $\gamma(x)$ is the same as $g(x)$ in (93), namely it is the half-wave device. Thus the quantity $y_+(t)$ (see diagrams) is given by

$$y_+(t) = \sum_{m=0}^{\infty} C(\nu, m) V^{\nu}(t) \cos [m(\omega_0 t + \phi(t))] \quad (99)$$

If we substitute $-x$ for x in (90) to evaluate $y_-(t)$, and use the fact that $I_m(-sV) = (-1)^m I_m(sV)$, then

$$y_-(t) = \sum_{m=0}^{\infty} C(\nu, m) (-1)^m V^{\nu}(t) \cos [m(\omega_0 t + \phi(t))] \quad (100)$$

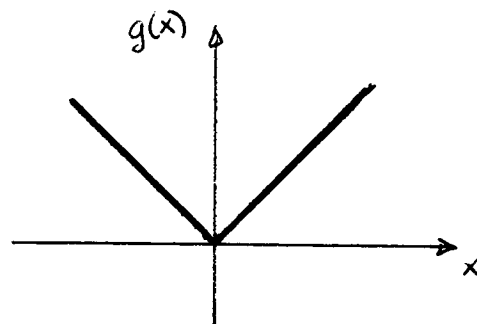
Thus for the $\begin{cases} \text{even} \\ \text{odd} \end{cases}$ full wave device:

$$y(t) = y_+(t) \pm y_-(t) = 2 \sum_{\substack{m=0 \\ m \begin{cases} \text{even} \\ \text{odd} \end{cases}}}^{\infty} C(\nu, m) V^{\nu}(t) \cos [m(\omega_0 t + \phi(t))] \quad (101)$$

Thus for an even full wave device, the output has harmonics which are even multiples of ω_0 . Odd multiples of ω_0 indicate the harmonics in the output of an odd law device.

Example 4: Suppose we wish to recover the envelope of $x(t) = V(t) \cos [\omega_0 t + \phi(t)]$

If we pass this through a full wave even device with $\nu = 1$, (see diagram below)



the output is given by

$$y(t) = 2 \sum_{\substack{m=0 \\ m \text{ even}}}^{\infty} C(1, m) V(t) \cos [m(\omega_0 t + \phi(t))]$$

If $V(t)$ and $e^{i\phi(t)}$ are slowly varying functions of time, then the Fourier transform of $V(t) \cos [m(\omega_0 t + \phi(t))]$ is non-zero for a small range

of f around $\pm \frac{m\omega_0}{2\pi}$, i.e., $V(t) \cos[m(\omega_0 t + \phi(t))]$ is a narrowband process about $\pm m f_0$. If we then pass $y(t)$ through a zonal filter $[f_0]$ (f), the output $Z(t)$ of the filter is given by

$$Z(t) = 2C(1,0)V(t) = \frac{2c}{\pi} V(t)$$

since all the harmonics for $m \neq 0$ are removed by the filter.

G. Random Variables

Not every voltage, electric field, etc. in a communication system is predictable, certainly not the information to be transmitted (otherwise why sent it?). The analysis of performance must be based on a thorough understanding of the random phenomena, which can occur in the process of communication. Rather than invoke all of the mathematical machinery of probability of statistics to handle this problem we shall describe here only a few of the basic ideas involved.

Suppose we desire to build a communication system to inform our alumnae of the number of girls in a 10 student engineering class. We know on the basis of previous experience that .95 of the classes contain no girls, .04 of the classes contain 1 girl and .01 of the classes (if we are lucky) contain 2 girls. No class ever contains more than 2 girls (a general principle). We can now define a function which describes our estimate of the future enrollment based on past experience.

n = number of girls in a class of 10 students

$P(n)$ = probability of n girls in a 10 student class

$$P(0) = .95$$

$$P(1) = .04$$

$$P(2) = .01$$

$$P(n) = 0, n \geq 3$$

(102)

$P(n)$ is called the probability function of n . Generally probability functions must satisfy the relations

$$\begin{array}{l} P(n) \geq 0 \\ \sum_n P(n) = 1 \end{array}$$

(103)

where N is the range of values of n . The number n is called a discrete random variable.

We may desire to compute the expected value of a random variable. For instance, in the above example a reasonable question to ask is, "What number of girls do you expect to be in the next engineering class?" To answer this question we compute the expected value of the random variable n .

$$E\{n\} = \sum_N nP(n) \quad (104)$$

In our example

$$E\{n\} = 0 \times .95 + 1 \times .04 + 2 \times .01 = .06$$

Thus the "average" number of girls is .06 in a class of 10. Notice that the expected value of a random variable is simply an average which is weighted by ^{our} a priori guess (before the fact) about the relative occurrence of the random variable.

Minor modifications are required to describe continuous random variables, i.e. random variables which can take on a continuum of values. Consider the problem of measuring the resistance of a 1Ω (nominal) resistor. Due to imperfections in the resistor, electronic noise in the meter, etc., the meter needle may indicate any of a continuum of values on its scale. Our model for this situation based on past experience might be the following:

$$r = \text{measured resistance in ohms}$$

$$p(r) = \frac{K e^{-\frac{(r-1)^2}{2(.01)}}}{\sqrt{2\pi}(.01)} U(r) \quad (105)$$

The function $p(r)$ is known as a probability density. To obtain the fraction of the time that we expect the needle to read between a and b , i.e. the "probability" that $a < r \leq b$, we need only compute the integral

$$P(a < r \leq b) = \int_a^b p(r) dr \quad (106)$$

This is a general characteristic of all probability densities. Along with (106) all probability densities must satisfy the conditions

$$\boxed{\begin{aligned} p(r) &\geq 0 \\ \int_R p(r) dr &= 1 \end{aligned}} \quad (107)$$

where R is the range of r .

The constant K in (105) is chosen to satisfy (107), in this case $K \equiv 1$. To calculate the expected value of r , we simply integrate r weighted by its probability density.

$$\boxed{E\{r\} = \int_R r p(r) dr} \quad (108)$$

Certainly there is an analogy between probability densities and mass densities. For example if we equate $p(r)$ with a mass density, then the mass located between a and b in a rod lying along the r axis is given by (106). Likewise (107) implies that negative mass cannot exist and that the total mass of the rod is unity. The expected value of r given in (108) is simply the center of mass of the rod. This analogy can be further extended. If we spin the rod about an axis perpendicular to the r axis, the spin axis passing through the center of mass, we have for the moment of inertia about $E\{r\}$:

$$\text{Var}\{r\} = E\{(r - E\{r\})^2\} = \int_R (r - E\{r\})^2 p(r) dr \quad (109)$$

The equivalent quantity to the moment of inertia about the center of mass is known as the variance of r . Notice that by expanding the right hand side of (109) we see that

$$\text{Var}\{r\} + E^2\{r\} = \int_R r^2 p(r) dr \quad (110)$$

which is the usual equation for the moment of inertia when the axis of rotation is translated. Several fundamental points must be understood: 1) $E\{r\}$ is not a function of r , but simply a number which depends on the shape of the probability density of r . 2) A function $f(r)$ of a

random variable r is a new random variable with a new probability density which can be derived from $p(r)$ and the function f . In deriving (109) we have used the consistency relation (which can be proven), namely

$$E\{f\} = \int_F f p(f) df = \int_R f(r) p(r) dr \quad (111)$$

Equation (111) states that the "probability mass" which you associate with a point f , is the same as the total mass which you associate with all points r which give the value $f(r)$ for the random variable. It is this idea which is used in the derivation of $p(f)$ from $p(r)$ and $f(r)$. In (109)

$f(r) = (r - E\{r\})^2$ and the expected value of this function of r is the variance of r .

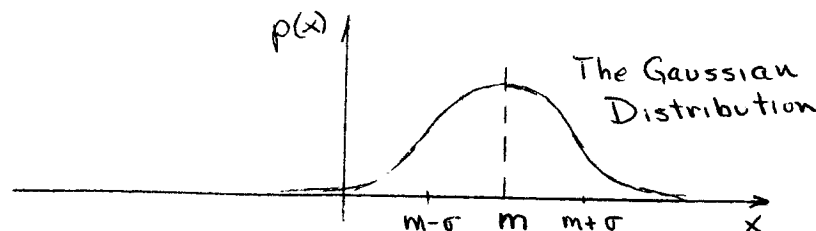
Example 5: A probability density which occurs repeatedly in communications is the Gaussian (or normal) density.

$$p(x) = \frac{e^{-\frac{1}{2} \left(\frac{x-m}{\sigma}\right)^2}}{\sqrt{2\pi} \sigma}, \quad -\infty < x < \infty \quad (112)$$

It can be verified that for the Gaussian random variable x ,

$$E\{x\} = m \quad (113)$$

$$\text{Var}\{x\} = \sigma^2 \quad (114)$$



We note that in the example of measuring the resistance of a 1Ω resistor, the probability density is Gaussian for positive values of r only. However, the parameter $m = 1$ and $\sigma = .1$ implies that for $r < 0$, (105) is essentially zero anyway and we might as well consider the mean variance of r to be 1 and .01 as in (112), (113) and (114).

The variance of a random variable is an indicator of the amount of uncertainty about the values which the random variable may take on. Suppose we calculate the probability that a random variable will be more than ϵ from its expected value. By (106):

$$P\{|x - E\{x\}| > \epsilon\} = \int_{-\infty}^{E\{x\} - \epsilon} p(x) dx + \int_{E\{x\} + \epsilon}^{\infty} p(x) dx \quad (115)$$

Continuing by placing a factor greater than 1 in each integrand and then expanding the region of integration,

$$P\{|x - E\{x\}| > \epsilon\} \leq \int_{-\infty}^{\infty} \left[\frac{x - E\{x\}}{\epsilon} \right]^2 p(x) dx = \frac{\sigma_x^2}{\epsilon^2} \quad (116)$$

Equation (116) is known as Tchebycheff's inequality. It indicates that for any positive number ϵ , as $\sigma_x^2 \rightarrow 0$, the probability x varies from its expected value by more than ϵ , goes to zero. Tchebycheff's inequality is the basis for many useful theorems in communication theory.

H. Random Processes

Interference or "noise" in a communication system is usually a continuous process, and as such, cannot be described by a single random variable. Suppose we imagine that we can write down or draw all the possible waveforms $x(t)$ which might occur in a system. If we sample the waveforms at a particular time, say t_0 , then $x(t_0)$ can be considered to be a continuous random variable having an appropriate probability density. Thus the ensemble of waveforms can be considered to be a (possibly uncountable) collection of random variables, a particular random variable being denoted by the time of the sample which it represents.

It is not enough to describe the process by the probability density of each of the random variables since their values are not generally determined independently. To describe the process completely, we must be able to state the joint probability density of any finite number of random variables in the random process. Suppose we now denote the joint probability density for

$x(t_1)$ and $x(t_2)$, as $p(x_1, x_2)$. The probability that in any waveform from the ensemble, $x(t_1)$ and $x(t_2)$ satisfy the relations

$$\begin{aligned} a_1 < x(t_1) \leq b_1 \\ a_2 < x(t_2) \leq b_2 \end{aligned} \quad (117)$$

is given by

$$P\{a_1 < x_1 \leq b_1, a_2 < x_2 \leq b_2\} = \int_{a_1}^{b_1} \int_{a_2}^{b_2} p(x_1, x_2) dx_2 dx_1 \quad (118)$$

Of course the joint density is subject to the conditions

$$\boxed{\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2) dx_1 dx_2 &= 1 \\ p(x_1, x_2) &\geq 0 \end{aligned}} \quad (119)$$

If we ask the question, "What is the probability that $-\infty < x_1 \leq b_1$ and $-\infty < x_2 < \infty$?", this is simply equivalent to asking "What is the probability that $-\infty < x_1 \leq b_1$?", since we allow x_2 to take on any value. Thus from (118) the joint probability density must satisfy the consistency relation

$$\int_{-\infty}^{b_1} \left[\int_{-\infty}^{\infty} p(x_1, x_2) dx_2 \right] dx_1 = \int_{-\infty}^{b_1} p(x_1) dx_1 \quad (120)$$

or differentiating with respect to the upper limit, we see that

$$\boxed{\int_{-\infty}^{\infty} p(x_1, x_2) dx_2 = p(x_1)} \quad (121)$$

In this manner we can obtain the density of $x(t_1)$ or $x(t_2)$ from the joint density $p(x_1, x_2)$.

Suppose that $t_2 > t_1$, and we observe $x(t_1)$ but we have not observed $x(t_2)$. We can ask the question, "What is the probability density of x_2 conditioned on the fact that x_1 is known?" A consistent method by

the axioms of probability, for calculating the conditional probability of $-\infty < x_2 \leq b_2$ for a given x_1 is to compute the probability of the simultaneous occurrence

$$b_1 - \Delta x < x_1 \leq b_1, -\infty < x_2 \leq b_2 \quad (122)$$

for very small values of Δx , and multiply by a suitable constant so that ^{the event $-\infty < x_2 < \infty$ given} the value of x_1 , has probability 1. We can do this quite easily via the following calculation:

$$\int_{-\infty}^{b_2} p(x_2|x_1) dx_2 = \lim_{\Delta x \rightarrow 0} \frac{Pr\{b_1 - \Delta x < x_1 \leq b_1, -\infty < x_2 \leq b_2\}}{Pr\{b_1 - \Delta x < x_1 \leq b_1, -\infty < x_2 < \infty\}} \quad (123)$$

Notice that when $b_2 = \infty$, the numerator of (123) is the same as the denominator and thus $p(x_2|x_1)$ integrates to 1 over $-\infty < x_2 < \infty$. Evaluating the right hand side of (123) with the aid of (118) and (121) gives

$$\begin{aligned} \int_{-\infty}^{b_2} p(x_2|x_1) dx_2 &= \lim_{\Delta x \rightarrow 0} \frac{\int_{-\infty}^{b_2} \left[\int_{b_1 - \Delta x}^{b_1} p(x_1, x_2) dx_1 \right] dx_2}{\int_{b_1 - \Delta x}^{b_1} p(x_1) dx_1} \\ &= \lim_{\Delta x \rightarrow 0} \frac{\int_{-\infty}^{b_2} p_{x_1, x_2}(b_1, x_2) \Delta x dx_2}{p_{x_1}(b_1) \Delta x} \quad (124) \end{aligned}$$

Cancelling Δx in numerator and denominator (we have already used the fact that Δx is small in deriving the second line of (124)) and noting that $x_1 = b_1$ in the limit as $\Delta x \rightarrow 0$, we have that

$$\int_{-\infty}^{b_2} p(x_2|x_1) dx_2 = \int_{-\infty}^{b_2} \frac{p(x_1, x_2) dx_2}{p(x_1)} \quad (125)$$

Differentiating with respect to b_2 , we have

$$p(x_2|x_1) = \frac{p(x_1, x_2)}{p(x_1)} \quad (126)$$

where $p(x_2|x_1)$ satisfies all the characteristics of a probability density. The density $p(x_2|x_1)$ is called the conditional probability density of x_2 given x_1 . If $p(x_2|x_1) = p(x_2)$ for all values of x_1 and x_2 , then our a priori conception of x_2 does not depend on the observed value of x_1 and we say that x_1 is independent of x_2 . If x_1 is independent of x_2 then (126) reduces to

$$p(x_1, x_2) = p(x_1)p(x_2) \quad (127)$$

The only case in which the joint density can be determined from individual densities is the case when the variables are independent.

We shall define the expected value of any function $g(x_1, x_2)$ of x_1 and x_2 as

$$E\{g(x_1, x_2)\} = \iint_{-\infty}^{\infty} g(x_1, x_2) p(x_1, x_2) dx_1 dx_2 \quad (128)$$

In particular, we define the correlation of x_1 and x_2 as

$$R_x(t_1, t_2) = E\{x_1 x_2^*\} = \iint_{-\infty}^{\infty} x_1 x_2^* p(x_1, x_2) dx_1 dx_2 \quad (129)$$

$R_x(t_1, t_2)$ is called the ensemble autocorrelation function of the random process $x(t)$.

Example 6: If the joint probability density of all finite numbers of samples of a random process is Gaussian, the random process is Gaussian. Two random variables from a Gaussian random process must therefore have a joint probability density of the form:

$$p(x_1, x_2) = \frac{1}{2\pi(1-\rho^2)^{1/2}\sigma_1\sigma_2} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-m_1}{\sigma_1}\right)^2 + \left(\frac{x_2-m_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-m_1}{\sigma_1}\right)\left(\frac{x_2-m_2}{\sigma_2}\right)\right]\right\} \quad (130)$$

where m_1 and m_2 are arbitrary real numbers, σ_1 and σ_2 are arbitrary positive numbers, and ρ is a real number, having absolute value less than 1.

Note that the right hand side of (130) can be factored as follows:

$$p(x_1, x_2) = \left\{ \frac{\exp\left[-\frac{1}{2}\left(\frac{x_1 - m_1}{\sigma_1}\right)^2\right]}{\sqrt{2\pi}\sigma_1} \right\} \left\{ \frac{\exp\left[-\frac{1}{2}\left(\frac{x_2 - [m_2 + \rho\sigma_2\left(\frac{x_1 - m_1}{\sigma_1}\right)]}{\sqrt{1-\rho^2}\sigma_2}\right)^2\right]}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_2} \right\} \quad (131)$$

We see that $p(x_1, x_2)$ is the product of two Gaussian functions, one with mean m_1 and variance σ_1^2 , and one with mean $m_2 + \rho\sigma_2\left(\frac{x_1 - m_1}{\sigma_1}\right)$ and variance $(1-\rho^2)\sigma_2^2$. If we integrate (131) with respect to x_2 , from (97) we should obtain the density of x_1 . Thus, using the fact that Gaussian functions must integrate to 1, we have

$$p(x_1) = \frac{\exp\left[-\frac{1}{2}\left(\frac{x_1 - m_1}{\sigma_1}\right)^2\right]}{\sqrt{2\pi}\sigma_1} \quad (132)$$

Applying (126) we see that the conditional probability density of x_2 given x_1 is

$$p(x_2|x_1) = \frac{\exp\left[-\frac{1}{2}\left(\frac{x_2 - [m_2 + \rho\sigma_2\left(\frac{x_1 - m_1}{\sigma_1}\right)]}{\sqrt{1-\rho^2}\sigma_2}\right)^2\right]}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_2} \quad (133)$$

We note further that if $\rho=0$, $p(x_2|x_1)$ is not a function of x_1 , and therefore x_2 is independent of x_1 .

With the aid of equations (112), (113), and (114) we can now calculate the correlation function of the random process in terms of the quantities m_1, m_2, ρ, σ_1 and σ_2 .

$$R_x(t_1, t_2) = E\{x_1 x_2\} = \int_{-\infty}^{\infty} x_1 p(x_1) \left[\int_{-\infty}^{\infty} x_2 p(x_2|x_1) dx_2 \right] dx_1 \quad (134)$$

By applying (112) and (113) the result of the inside integration is the mean of the probability density of x_2 given x_1 .

$$\begin{aligned}
 R_x(t_1, t_2) &= \int_{-\infty}^{\infty} x_1 \left[m_2 + \rho \sigma_2 \left(\frac{x_1 - m_1}{\sigma_1} \right) \right] p(x_1) dx_1 \\
 &= (m_2 - m_1 \rho \frac{\sigma_2}{\sigma_1}) \int_{-\infty}^{\infty} x_1 p(x_1) dx_1 + \rho \frac{\sigma_2}{\sigma_1} \int_{-\infty}^{\infty} x_1^2 p(x_1) dx_1
 \end{aligned} \quad (135)$$

The first of these two integrals can be solved again using (112) and (113), and the second can be solved with the aid of (110), (112) and (114).

$$\begin{aligned}
 R_x(t_1, t_2) &= (m_2 - m_1 \rho \frac{\sigma_2}{\sigma_1}) m_1 + \rho \frac{\sigma_2}{\sigma_1} (\sigma_1^2 + m_1^2) \\
 &= m_1 m_2 + \rho \sigma_1 \sigma_2
 \end{aligned} \quad (136)$$

The function ρ depends on the choice of sample times t_1 and t_2 for the calculation of correlation, and is usually called the normalized covariance of the process.

$$\rho(t_1, t_2) = \frac{R_x(t_1, t_2) - m_1 m_2}{\sigma_1 \sigma_2} \quad (137)$$

1. Spectral Densities of Stationary Processes

As we already know, transient analysis of circuits with deterministic signals is more difficult than steady-state analysis under the same conditions. The same effect is present when we consider the input of a circuit to be a random process. The term analogous to steady-state, which applies to random processes is the word stationary. There are several ways of picturing stationary processes: a) If one of an ensemble of time functions will occur with probability P , then one of the same ensemble translated by an arbitrary amount of time, will also occur with probability P ; or b) the expected value of any function of the random process, is independent of time, or c):

Definition: A random process is stationary if for every arbitrarily selected finite number of samples of the process, and every value of the parameter T , the probability density of the samples satisfies the relation

$$p(x(t_1), x(t_2), \dots, x(t_n)) = p(x(t_1+T), x(t_2+T), \dots, x(t_n+T))$$

Thus, the probability density of two samples of a random process $p(x(t), x(t+\tau))$ can only be a function of the time difference τ ^(and not t) if the process is stationary. Furthermore, from an inspection of (129) with $t_1 = t+\tau$, $t_2 = t$, we see that the integrand of (129) is only a function of τ and we can re-define the correlation function of a stationary process to be

$$R_x(t+\tau, t) \triangleq R_x(\tau) \quad (138)$$

In the analysis of linear systems, we found it convenient to transform the input of the system and deal with a frequency representation of the output. Unfortunately, sample functions of stationary random processes are not generally Fourier transformable, and even if they were, we must deal with a whole ensemble of possible input processes, instead of one input signal. Recall however, in our discussion of deterministic periodic signals, an alternate derivation of the spectral density of a periodic signal was given by the Fourier transform of the time autocorrelation function. For stationary processes, in most cases of interest we have

$$R_x(\tau) = E\{x(t+\tau)x^*(t)\} = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t+\tau)x^*(t) dt = R_x(\tau) \quad (139)$$

That is, the result of calculating the ensemble autocorrelation function $R_x(\tau)$ as an average of $x(t+\tau)x^*(t)$ for any given t is equivalent to calculating the time average of $x(t+\tau)x^*(t)$ for one sample function of the process. Processes which satisfy (139) are said to be correlation ergodic. Our intuition should now tell us that a good representation of the spectral density of a stationary process is the Fourier transform of its ensemble autocorrelation function.

Definition: The spectral density of a stationary random process $x(t)$ is given by:

$$S_x(f) = \int_{-\infty}^{\infty} R_x(\tau) e^{-i\omega\tau} d\tau \quad (140)$$

The average power in $x(t)$ is given by $E\{|x(t)|^2\}$.

The average power in a random process can be reduced to

$$E\{|x(t)|^2\} = R_x(0) = \int_{-\infty}^{\infty} S_x(f) df \quad (141)$$

by use of the definition of a correlation function and the Fourier transform relation between $R_x(\tau)$ and $S_x(f)$.

The spectral density of a real random process can also be shown to have the properties

$$\begin{aligned} S_x(f) &\geq 0 \\ S_x(f) &= S_x(-f) \end{aligned} \quad (142)$$

Example 7: Consider the random process $x(t) = \cos(\omega_0 t + \theta)$ where ω_0 is known and θ is a random variable with $p(\theta) = \frac{1}{2\pi} 1_{2\pi}(\theta)$

The time correlation function of one sample function $\cos(\omega_0 t + \theta_0)$ is given by

$$\begin{aligned} R_x(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \cos(\omega_0 t + \omega_0 \tau + \theta_0) \cos(\omega_0 t + \theta_0) dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \frac{1}{2} [\cos \omega_0 \tau + \cos(2\omega_0 t + \omega_0 \tau + 2\theta_0)] dt \\ &= \frac{\cos \omega_0 \tau}{2} \end{aligned} \quad (143)$$

The ensemble correlation function of the process is given by (using (111)):

$$\begin{aligned} R_x(t+\tau, t) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\omega_0 t + \omega_0 \tau + \theta) \cos(\omega_0 t + \theta) d\theta \\ &= \frac{\cos \omega_0 \tau}{2} \triangleq R_x(\tau) \end{aligned}$$

(144)

From (143) and (144) we have verified that $x(t)$ is correlation ergodic. It can be verified that $x(t)$ is actually stationary. The spectral density of the process is (using (46)):

$$S_x(f) = \frac{1}{4} [\delta(f-f_0) + \delta(f+f_0)] \quad (145)$$

and the average power is $R_x(0) = \frac{1}{2}$.

It is now possible to investigate the basic properties of the output $y(t)$ of a linear filter when the input $x(t)$ is a stationary random process. We can derive the average value of the output as follows:

$$m_y(t) = E \left\{ \int_{-\infty}^{\infty} x(t-\alpha) h(\alpha) d\alpha \right\} \quad (146)$$

where $h(t)$ is the impulse response of the linear filter. In (146) we have used the convolution integral representation for the output of a linear filter. In almost all cases of practical interest it is possible to interchange the order of integration over α and computation of expected value, since both are linear operations. If we do this, then (146) reduces to

$$m_y(t) = \int_{-\infty}^{\infty} E \{ x(t-\alpha) h(\alpha) \} d\alpha \quad (147)$$

Since for fixed values of t and α , $x(t-\alpha)$ is a random variable with known expected value $m_x(t-\alpha)$ and $h(\alpha)$ is a constant, we have that

$$m_y(t) = \int_{-\infty}^{\infty} m_x(t-\alpha) h(\alpha) d\alpha \quad (148)$$

Thus we can treat the expected value of a random process as a deterministic signal in analyzing linear operations.

The output ensemble (or time) autocorrelation function can now be computed using the same techniques.

$$\begin{aligned} R_y(t+\tau, t) &= E \{ y(t+\tau) y^*(t) \} \\ &= E \left\{ \left[\int_{-\infty}^{\infty} x(t+\tau-\alpha) h(\alpha) d\alpha \right] \left[\int_{-\infty}^{\infty} x(t-\beta) h(\beta) d\beta \right]^* \right\} \quad (149) \end{aligned}$$

Using the fact that conjugating an integral is equivalent to conjugating the integrand, and computing expected values before integrating over α and β gives

$$\begin{aligned} R_y(t+\tau, t) &= \iint_{-\infty}^{\infty} E\{x(t+\tau-\alpha)x^*(t-\beta)\}h(\alpha)h^*(\beta)d\alpha d\beta \\ &= \iint_{-\infty}^{\infty} R_x(t+\tau-\alpha, t-\beta)h(\alpha)h^*(\beta)d\alpha d\beta \end{aligned} \quad (150)$$

Let us now assume that $x(t)$ is stationary and therefore that the correlation function of $x(t)$ depends only on the difference in sample times. Thus, using (138):

$$R_y(\tau) = \iint_{-\infty}^{\infty} R_x(\tau-\alpha+\beta)h(\alpha)h^*(\beta)d\alpha d\beta \quad (151)$$

This expression simplifies greatly if we take Fourier transforms of both sides. Then

$$S_y(f) = \int_{-\infty}^{\infty} e^{-j\omega\tau} \left[\iint_{-\infty}^{\infty} R_x(\tau-\alpha+\beta)h(\alpha)h^*(\beta)d\alpha d\beta \right] d\tau \quad (152)$$

Let $\tau' = \tau - \alpha + \beta$ Then the integrals can be separated.

$$S_y(f) = \int_{-\infty}^{\infty} e^{-j\omega\tau'} R_x(\tau') d\tau' \int_{-\infty}^{\infty} e^{-j\omega\alpha} h(\alpha) d\alpha \int_{-\infty}^{\infty} e^{j\omega\beta} h(\beta) d\beta$$

$$\boxed{S_y(f) = S_x(f) |H(f)|^2}$$

(153)

By using the transform domain, we have been able to significantly simplify the computation of the statistical characteristics of the output of a linear system.

Example 8: Let $x(t)$ be a real, stationary, Gaussian random process with mean zero and correlation function $\sigma^2 e^{-|\tau|}$. Note that the average power in $x(t)$ is given by (using (86) and (117)):

$$\text{Var} \{x(t)\} = E \{ |x(t)|^2 \} = R_x(0) = \sigma^2 \quad (154)$$

and the normalized covariance of $x(t)$ is given by

$$\rho_x(\tau) = e^{-|\tau|} \quad (155)$$

The spectral density of $x(t)$ is given by Fourier transform pair (49).

$$S_x(f) = \sigma^2 \frac{2}{1 + \omega^2} \quad (156)$$

Let $x(t)$ be the input to an R-C filter with impulse response $\frac{1}{RC} U(t) \exp\{-t/RC\}$ and hence system function $[1 + (RC)(i\omega)]^{-1}$ (See example 3).

We can use the basic result (153) to find the spectral density of the filter output $y(t)$.

$$\begin{aligned} S_y(f) &= \sigma^2 \left(\frac{2}{1 + \omega^2} \right) \left| \frac{1}{1 + (RC)(i\omega)} \right|^2 \\ &= 2\sigma^2 \left(\frac{1}{1 + \omega^2} \right) \left(\frac{1}{1 + (RC\omega)^2} \right) \end{aligned} \quad (157)$$

To find the output ensemble correlation function using the transform table, we can make a partial fraction expansion of (157).

$$S_y(f) = \frac{2\sigma^2}{1 - (RC)^2} \left[\frac{1}{1 + \omega^2} - \frac{(RC)^2}{1 + (RC)^2 \omega^2} \right] \quad (158)$$

Using (49) to calculate the correlation function of $y(t)$ gives

$$R_y(\tau) = \frac{\sigma^2}{1 - (RC)^2} \left[e^{-|\tau|} - RC e^{-|\tau|/RC} \right] \quad (159)$$

Since the expected value of the input is zero, the expected value of the output is also zero by (148). Thus the average power in the output and hence the variance of the output are given by:

$$\text{Var} \{y(t)\} = R_y(0) = \frac{\sigma^2}{1 + RC} \quad (160)$$

One very important result which we shall not prove here is that linear operations on Gaussian processes yield Gaussian processes. Thus from (159) and (160) we see that the normalized covariance of $y(t)$ is $(e^{-|t|} - R e^{-|t|/R}) / (1 - R)$. The joint density of $y(t)$, $y(t + \tau)$ is of the same form as (130), with appropriate substitutions.

PART II

COMMUNICATION SYSTEMS

A. Noise Sources

The radio communication channel is characterized by the type of interference that it places on the reception of electromagnetic energy radiated from the transmitter. These disturbances may be divided into three categories. One always present, and strictly unavoidable form of interference, is thermal noise in the receiver components. Modern technological advances in low temperature receivers, however have reduced this form of interference by an order of magnitude or more. In high frequency communication via the ionosphere and in channels employing tropospheric propagation, often a more serious form of interference is fading and multipath propagation of the signals. This can usually be characterized as a random linear time varying transformation on the transmitted signal. The third type of interference is man-made electromagnetic radiation at frequencies within the receiver band. This is of least interest in all but certain military applications (for example: jamming) since it can be avoided by providing regulatory and logistic precautions.

Since there has been an increasing emphasis on line-of-sight communication with space vehicles and satellite relay systems, the study of communication channels perturbed only by additive thermal noise has extensive significance. Since this is the one unavoidable form of radio frequency interference common to all space communication applications, it is reasonable that it be considered first. Furthermore, the performance analysis of some of the analog and digital communication systems we shall consider are relatively simple when the disturbances are additive. Therefore, we shall restrict attention solely to communication in the presence of additive thermal noise.

The thermal noise power spectral density in an R ohm resistor is

$$2 k T R$$

where $k = 1.38 \times 10^{-16}$ ergs/degree is Boltzmann's constant, and T is the temperature in degrees Kelvin. We shall refer to the one-sided noise power density normalized to a one ohm resistor as

$$N_o = 4 k T \text{ watts/cps}$$

Thus, if the receiver system temperature is T degrees Kelvin, and the one-sided noise bandwidth is W cps, the total noise power at the receiver front end is $N_o W$ watts.

B. Received Signal Power

We will also find it necessary to know the power arriving at the receiver which was radiated from the transmitter; that is, signal power at the receiver. This will be determined in terms of certain significant parameters which characterize the transmitter and receiver subsystems.

Towards this purpose, an "effective area" of the receiving antenna, A_r , is defined so that the useful power, P_r , received by the receiving antenna is given by this area multiplied by the average power density, D_r , in the oncoming wave. Hence

$$P_r = A_r D_r$$

For antennas with large apertures and with uniform illumination, the effective area is approximately equal to the aperture area.

Many antennas have a certain directivity as compared with an imagined isotropic radiator, which radiates (and receives) equally in all directions. This is an advantage if we know the direction from which the radiation is arriving, since it results in considerable savings in power. The amount of savings is frequently expressed as the gain, g , of the antenna, defined as the ratio of power required from the isotropic radiator to produce the given intensity in the desired direction to that required from the actual antenna. Antenna gain is in general a function of direction about the antenna, and of the condition of impedance match in the associated wave guide. When not otherwise specified, it will be assumed to be the value for a matched load and for the direction of maximum gain.

The power density at the receiver is the directional power density at the transmitter, D_t , divided by $4\pi R^2$, where R is the distance from the transmitting to the receiving antenna. The surface area of a sphere of radius R is $4\pi R^2$

Thus

$$D_r = \frac{D_t}{4\pi R^2}$$

In terms of the transmitter power, P_t , the transmitter power density is

$$D_t = g_t P_t$$

Combining these relationships, the total received power is

$$P_r = \frac{A_r g_t P_t}{4 \pi R^2}$$

This is one form of what is often termed the radar equation.

Therefore, with knowledge of P_r and N_o , we can specify a value for signal-to-noise power ratio, which we shall see is a fundamental system parameter.

C. Amplifiers

It has been noted that noise is an unavoidable part of any communication system. In space telemetry systems, this noise is essentially additive, white (i.e. flat spectral density) and Gaussianly distributed. Since, as we shall see, the ratio of the signal power to the noise power essentially determines the performance of any communication system, an improvement can be attained by either increasing the signal power at the receiver, or decreasing the noise power. We shall now examine methods for accomplishing both of these tasks.

The most significant noise contribution in most space communication environments arises in the initial stages of the receiver. Since the transmitter operates at relatively high signal levels, the signal-to-noise ratio at the transmitter can be kept very large. Background radiation at the radio frequencies generally used is relatively insignificant (and, in any event, unavoidable). At the receiver, however, the signal power is extremely low so that any noise contributed in the initial process of amplifying this signal may be, in comparison, most significant. It is at this stage that the greatest effort is demanded to decrease the additive noise, and it is here that the most important progress has been made.

The earliest technique for the low-noise amplification of microwave frequencies involved the use of the "traveling-wave tube." The traveling wave tube developed during World War II relies upon the interaction between an electron beam by passing it through a wave guide, generally in the

shape of a helix. Since electromagnetic energy traverses linearly along a waveguide at nearly the velocity of light c , its rate of progress along the axis of the helix is approximately $(h/L)c$, where h is the length of the axis of the helix, and L is the length of the wave guide comprising the helix. Thus it is possible to make the velocities of linear propagation of the signal and the electron beam equal. When this is done, there is an interaction between the electric field of the signal and the electrons in the beam. The electron density is increased or decreased depending upon the intensity and direction of the field. This "bunching" in turn causes the field to be intensified in proportion to its original strength, thus producing amplification. Extremely large amplifications over a wide band of microwave frequencies are, indeed, possible with this technique. The noise arises, as usual, because the electrons do not all have the same energy or velocity. Thus the bunching cannot be perfect. Since the electrons are not all moving with the same velocity, they exhibit a counter-tendency toward a random distribution. This appears as noise at the output. Much effort has been made to decrease the noise inherent in traveling wave tubes, and amplifiers using these tubes have been built with effective noise temperatures of less than 300 K.

We now consider the two more recent and now principal types of low noise amplifiers: the parametric amplifier and the solid state maser.

D. Parametric Amplifier

The action of the parametric amplifier is commonly compared to the method in which a child, sitting in a swing, is able to increase the amplitude of his swinging arc. At the height his displacement, when the swing changes directions, the child pulls back on the ropes, thereby slightly increasing his height, and hence his potential energy. At the bottom of the arc, the tension on the ropes is relaxed so that this potential energy is entirely converted into kinetic energy. Because the maximum height was increased this kinetic energy is greater than it would have been and, at the next peak, the potential energy has increased over its value at the previous peak. The energy of the child, therefore, is converted into oscillation energy of the swing.

In the same way any oscillator or resonant device can gain energy by being "pumped" at the right times. In fact, it can be shown that the oscillator exhibits a net energy gain even if it is pumped at the wrong time, too, so long as at least some of the pumping occurs at the peaks and troughs of the potential and kinetic energy storage cycles.

As suggested by the name, a parametric amplifier (or paramp) consists of a device in which a parameter, such as an inductance L or a capacitance C , is varied periodically. Recalling the energy relations

$$W_1 = \frac{1}{2} L i^2$$

$$W_2 = \frac{1}{2} C v^2$$

where i is the current through L and v is the voltage across C , it is seen that varying L or C can vary the energy in the circuit. Moreover, it is noted that the phase relation between the varying parameter and the signal is important for efficient action. In the case of a variable capacitance it is noted that the energy can be increased by decreasing C when v is near the positive or negative maxima (since $v = Q/C$, Q being the charge). In practice, a resonant circuit, tuned to the signal frequency, is invariably used because it is difficult to obtain large changes in capacitance.

To illustrate the mechanism by which energy can be transferred from a "pump" which drives an energy storage element into the fields of a resonant tank, consider the simple resonant circuit of Fig. 4-1. Imagine that it is possible to pull the capacitor plates apart and push them together again at will. Suppose that at the time the voltage across the capacitor goes through a positive or negative maximum the plates are suddenly pulled apart. Work must be done in separating the charge on the two plates. The energy goes into the electric fields existing across the plates. The capacitance is reduced and, using $v = Q/C$, the voltage is amplified. Each time the voltage goes through zero, the plates are suddenly pushed back together again. When the plates are pushed together there is no charge on the capacitor, so no work is done or required in this operation. The net result is amplification of the voltage across the capacitor, the flow of energy being from whatever pumps the plates to the fields of resonant circuit. This is illustrated in the voltage curve labeled (a).

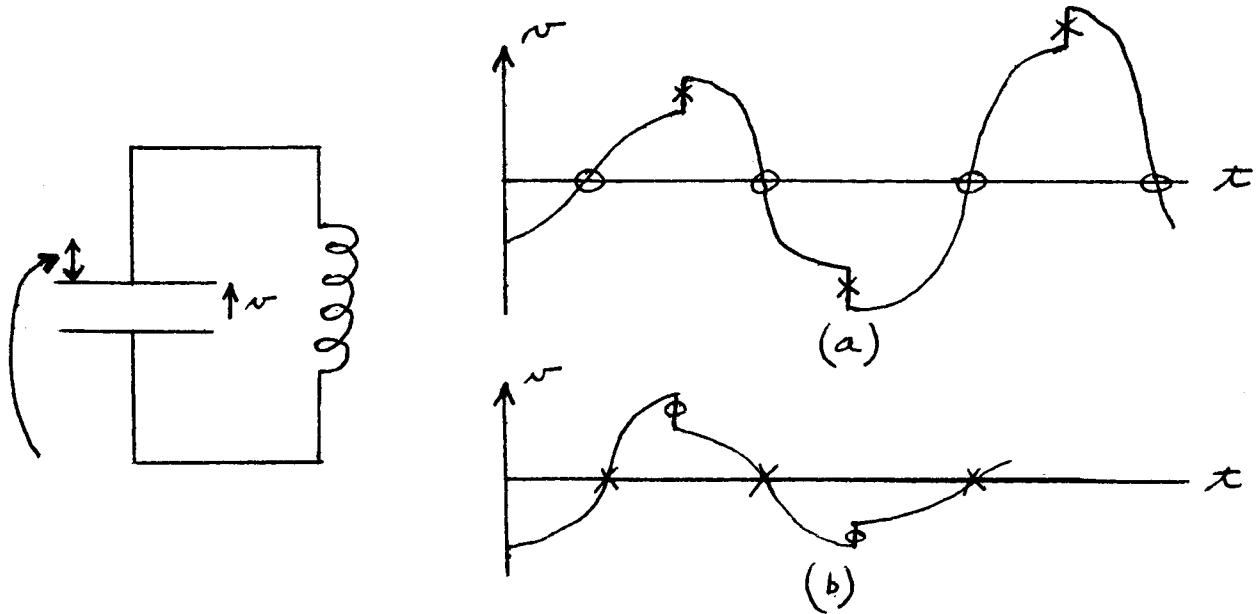


Fig. 4.1 Simple Resonant Circuit and pump to illustrate the mechanism for the variable parameter principle.

The crosses indicate a sudden pulling apart of the plates, the circles a sudden pushing together of the plates. Note that for this circuit, the pumping is repeated periodically at twice the frequency of the signal. Note also that here a phase condition is necessary for amplifying the circuit voltage. If the plates are pushed together when the voltage is high and pulled apart when it is zero, the energy flow is in the opposite direction. The voltage is then attenuated as shown in curve (b).

The next step in the description of a parametric amplifier is to consider a two-tank circuit as in Fig. 4-2. Here, there is no phase restriction. In this case, the variable capacitance serves to couple together two different tank circuits of resonant frequencies ω_1 and ω_2 , respectively. The variable capacitor is driven sinusoidally at a rate $\omega_3 = \omega_1 + \omega_2$. If a voltage exists across one of the tanks at its resonant frequency, a second voltage is developed across the second tank at its resonant frequency by the mixing action in the variable capacitance. The phase of the second voltage is automatically adjusted so that the net energy flows into the tank circuits from the pumped capacitor.

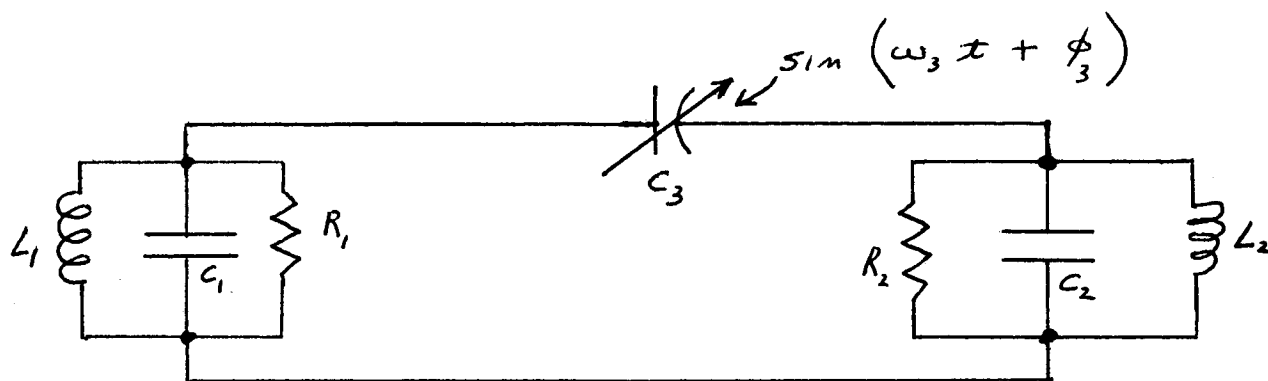


Fig. 4.2 An equivalent circuit for the two tank variable parameter system.

Thus far, what has been described is obviously suitable for setting up and maintaining oscillations. The energy transfer mechanism is also suitable for amplification. For example, in Fig. 4-2, assume we couple a signal generator and an output load into one of the tanks, say tank 1. The signal generator, of course, should be tuned as closely as possible to ω_1 . The magnitude of the capacitor variation, that is C_3 , should be reduced to a value just below the point where oscillations occur. An amplified version of the input signal will then appear at the output load.

What has been described here is a two-terminal device. An immediate problem arising from such devices is that there are no separate input and output terminals as in a vacuum tube amplifier, for example. Fortunately, this problem was solved by the concurrent development of the ferrite circulator. This device permits a separation of incident and reflected waves and thus provides the equivalent of input and output terminals. Fig. 4-3 shows a schematic diagram of an antenna coupled to a load through a two-terminal amplifier. Without a circulator it would be difficult to exploit the low noise properties of a negative resistance amplifier such as the paramp because the noise power from the load resistor would be amplified along with the incident signal and noise. What actually happens to the noise power from the load resistor is that it is radiated out into space by the antenna.

The parametric amplifier is a low noise device principally because the shot noise inherent in vacuum tube amplifiers is absent. The salient feature of these devices is that no refrigeration is required, thus

making its cost only a small fraction of that required for the maser, which requires cooling to liquid nitrogen temperature. It is generally conceded, however, that the maser is a superior device, and a discussion of the maser follows.

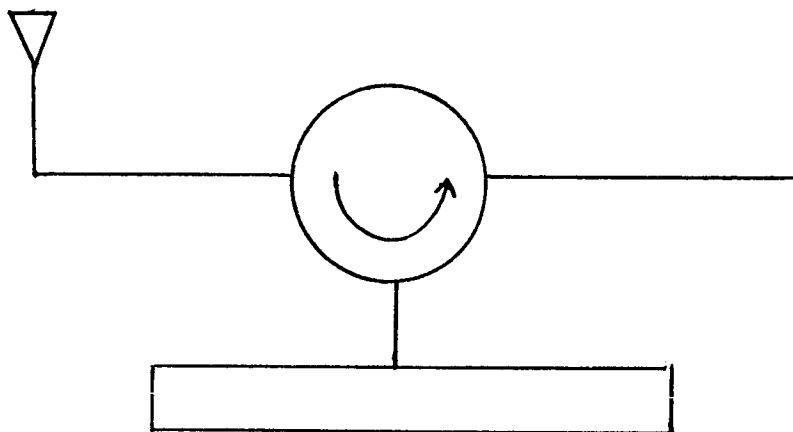


Fig. 4.3 Utilization of circulator with two terminal parametric amplifiers

E. Masers

Probably the most successful low-noise amplifier yet developed, however, is the "maser" (acronym for microwave amplification by simulated emission of radiation). The electrons in the crystal lattice of any material, like all electrons, spin about some axis. The orientation of the spin axes are restricted to certain positions, and normally the vast majority of the electrons are in the lowest energy position. If the difference in energy between the lowest two energy levels is ΔE , an amount of energy ΔE is absorbed by the crystals when an electron makes a transition from the lowest to the next lowest level, and an amount of energy ΔE is radiated when the reverse transition occurs. Normally transitions occur equally often in each direction so that the net radiated energy is zero. The frequency of this radiation, we have Plank's equation, must be

$$f = \frac{\Delta E}{h}$$

where $h = 6.6 \times 10^{-34}$ joules seconds. If the crystal is radiated with

energy at the frequency $\Delta E/h$, electrons are caused to make the transition to the higher level and energy is absorbed. By irradiation with energy at a higher frequency f' it is possible to excite the electrons to a still higher energy level $E' = hf'$. By the proper selection of a crystal it is possible to achieve a situation in which electrons, excited to the level $\Delta E'$ can decay to the level ΔE , but cannot decay further, to the ground level except in the presence of external radiation at the frequency $f = \Delta E/h$. It is thus possible to create a situation in which the majority of electrons are at the next to the lowest energy level. When this is the case, a signal at the frequency f when applied to the crystal exhibits a net increase in energy due to the preponderance of electron transitions to the ground level which it triggers. Thus, energy is transferred from the higher excitation frequency to the crystal, and from the crystal to the lower signal frequency. Again, the resulting amplification can be sizeable.

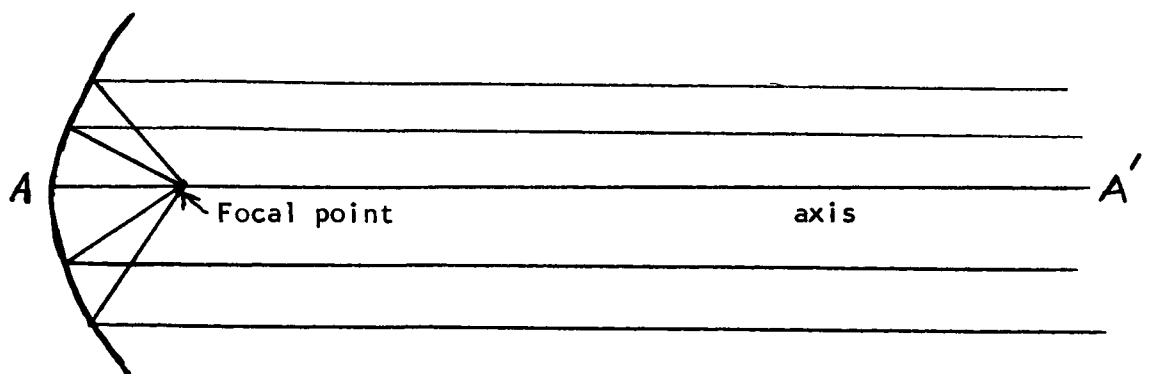
The noise generated in a maser amplifier can be exceedingly small. It is due to fluctuations in the radiation field in the neighborhood of the crystal. These fluctuations can be caused by thermal agitation of the electrons causing a noise spectral density $N_{th} = kT$ where T is the actual temperature (in degrees Kelvin) of the crystal. In addition, however, radiation is emitted due to spontaneous electron transitions which according to quantum mechanics, give rise to noise with a spectral density $N_q = \frac{1}{2}hf$. For microwave frequencies the total noise $N_o = N_{th} + N_q$ can be quite small and masers have been built with an effective noise temperature of less than $10^0 K$. Note, however, that the N_q term is independent of temperature and hence cannot be reduced by cooling the crystal. This term, being proportional to frequency, becomes more significant at higher frequencies. At frequencies in the visible light range, for example, (masers which operate at frequencies approximating those of visible light are called lasers, the m of microwave becoming the l of light) the effective noise temperature increases to about $20,000^0$ to $30,000^0 K$, thus seriously countering some of the real advantages associated with the use of lasers in space communications.

F. Antennas

Another method for overcoming the severe conditions encountered in space communications due to the vastly increased distances between the transmitter and receiver is through the use of high gain antennas. Since we are interested in communication from a point to a point rather than from a point to many points, as in commercial radio, we clearly want a somewhat different antenna design than that commonly used in the latter case.

In conventional AM radio transmission it is desired to radiate equally in all horizontal directions. To accomplish this, vertical or "dipole antennas" are used with heights which are ideally, one half of the wave length of the frequency radiated. Since they radiate horizontally, with little energy being transmitted vertically, they exhibit gains which are greater than one; in the case of an ideal dipole antenna the gain in the equatorial plane is 2.15db.

For space communication, however, it is desired to radiate energy only between the one transmitter and the one receiver. (There may be more than one receiver in practice, but at deep space distances, the earth itself is effectively one point.) It is therefore necessary to be able to direct the radiated energy in a narrow beam towards the receiver. This is most effectively accomplished by focusing the energy by means of a reflector in the shape of a paraboloid. A parabola, it will be recalled, has the geometric property illustrated in Fig. 6-1. To the extent that the angle of incidence of a microwave beam is equal to the angle of reflection, all energy originating at the focal point and striking the antenna, will be reflected in a direction parallel to the axis of the antenna A-A'.



6.1 Direction of reflection from a parabolic antenna.

Unfortunately, however, the wavefront will not remain constant with a diameter equal to that of the antenna, but will increase in area. For an intuitive understanding of the reason for this spread consider the illustration in Fig. 6-2. The observer at point C "sees" energy reflected from various portions of the surface of the antenna. Consider the signal reflection points A and B. Since C is closer to A than to B the energy from B must travel farther before it reaches the point C. If the geometry is such that the distance BC is exactly one-half wave length further than the distance AC, then the radiation from the two points A and B will arrive at C exactly 180° out-of-phase with respect to each other. The electromagnetic fields will have equal amplitudes but opposite signs and will, therefore, completely cancel each other. When C is too close to the axis, there will be no two points on the surface of the antenna such that the difference in their distance to C is as great as one-half wave length. As the distance from C to the axis increases, the points A and B satisfying the property described above will move closer together, and in addition, other points A' and B' can be found on the surface of the antenna such that the distance A'C and B'C differ by exactly three-halves wave length.

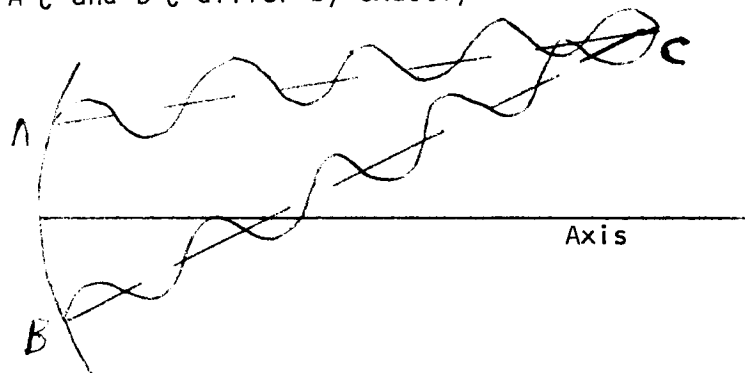


Fig. 6.2 Interference due to reflections from different parts of the Antenna.

Thus as C moves away from the axis, there is more and more cancellation, so that the net amount of energy striking C rapidly diminishes.

To theoretically determine the width of the beam at a distance R from the antenna consider the diagram in Fig. 6.3. We want to find a measure of the beam width, the smallest value of r , the distance from C to the axis, such that there is total cancellation of energy arriving from at least two points on the antenna. Clearly, the first two points

on the antenna surface which provide such cancellation are those two points in the same plane as C and separated by the maximum distance. Therefore, let A be at one extreme of the antenna and B at the other, separated by a distance d , the diameter of the antenna. Assuming that R and r are large compared to the dimensions of the antenna, and that the wavelength, λ , is small compared to d , it is easy to determine the value of r in terms of wavelength, the antenna diameter, and the distance R . First we find the point B' on the line CB such that the distances CA and CB' are equal. Since R is large, ϕ' is nearly a right angle and hence $\phi'_1 \approx \phi_1$. Clearly, $\phi'_2 = \phi_2$ and hence, the triangles $BC'C$ and $AB'B$ are (nearly) similar. Thus

$$\frac{AB'}{BB'} = \frac{BC'}{CC'}$$

and since

$$AB' = \sqrt{d^2 + \left(\frac{\lambda}{2}\right)^2} \approx d$$

and

$$BB' = \frac{\lambda}{2}$$

in order that we obtain the desired cancellation,

$$BC' = R, \quad \text{and} \quad CC' = r + \frac{d}{2} \approx r$$

we have after substitution

$$r \approx \frac{R\lambda}{2d}$$

and the beam width is proportional to $R\lambda/d$.

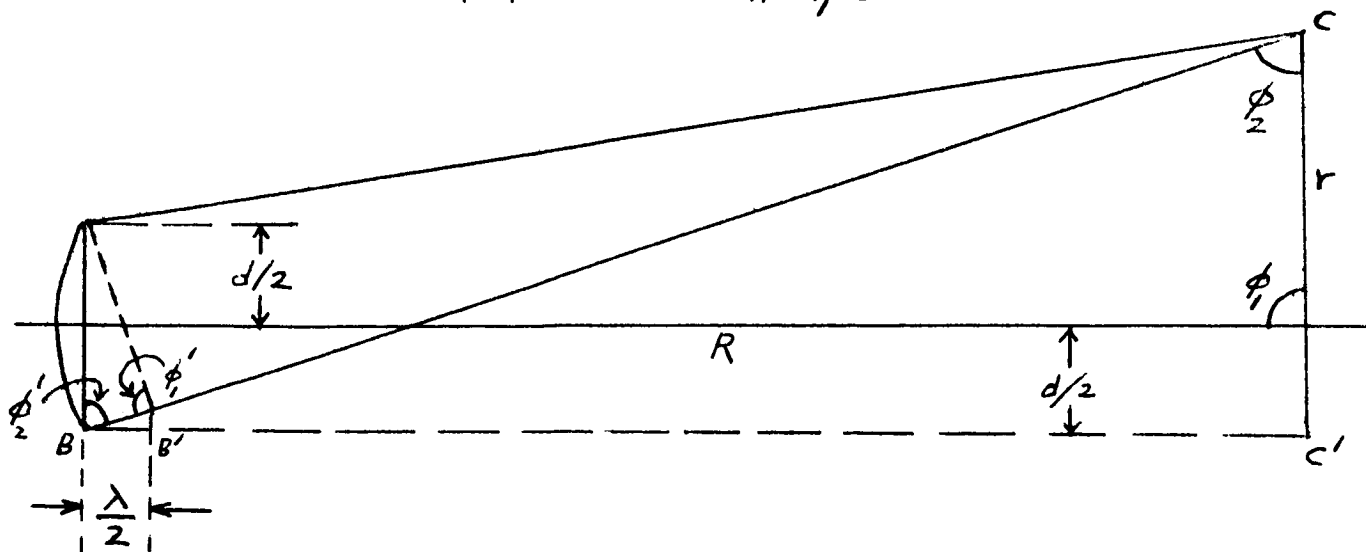


Fig. 6.3 Determination of the beam width.

This also demonstrates the important fact that as the diameter of the antenna is increased, the beam width is made smaller.

Now we again consider the power received by a second parabolic antenna of area A_r at a distance R from the first. Since the beam width is proportional to $2r$ its area is proportional to $(2r)^2$ and hence to $R^2 \lambda^2 / d^2$. The percentage of the power which is received, assuming $A_r < R^2 \lambda^2 / d^2$ is clearly proportional to the ratio of the area of the receiving surface to the area of the beam, since all the power striking the antenna surface is reflected to the focal point (assuming again a parabolic receiving antenna) and hence to the receiver input. Consequently, designating by P_t the total transmitted power, and P_r the total received power, we have

$$P_r \propto P_t \frac{A_r}{R^2 \lambda^2 / d^2}$$

And finally, since the area of the transmitter antenna is proportional to d^2 we have

$$P_r \propto P_t \frac{A_r A_t}{\lambda^2 R^2}$$

Actually, this heuristically derived result can be shown to be exact, so that

$$P_r = P_t \frac{A_r A_t}{\lambda^2 R^2}$$

This is a second form of the radar equation discussed previously.

As previously indicated, for non-ideal parabolic antennas, A_r and A_t must be replaced by an effective area which always is somewhat less than the true area. This is primarily because of the fact that it is difficult to radiate the entire surface of the antenna with energy of equal magnitude and equal phase. Typically, the effective area is 50% to 80% of the actual area.

In order to maximize the amount of power received, or equivalently, to maximize the gains of the two antennas, it is necessary to make the parabolic antennas as large as possible and the wavelengths as short as possible. First of all, there are practical limitations to the shortness of the wavelength. One of these limitations stems from the fact that, beyond a certain point, the effective noise temperature of the best amplifiers increases sharply as the wavelength decreases, thereby counteracting

the advantages in antenna gain. In addition, in order to realize the theoretical gains of parabolic antennas, the dimensions of the antenna must be accurate to within a fraction of a wavelength. Since the gain increases in proportion to the area, it is advantageous to make the area as large as possible. But the larger the area, the more difficult it is to keep the tolerances within the necessary limitations. Thus, there is a trade-off between the area and the wavelength. Moreover, because the transmitting and receiving antennas are moving with respect to one another it must be possible to move the ground based antenna so that this also places restrictions on its size, (the fact that the vehicle antenna must be propelled through space, of course, limits its size). Finally, the transmitter antenna must be pointed in space with an accuracy proportional to the width of the beam or the maximum energy is not received at the receiver antenna. This clearly also limits the gain, and becomes particularly significant at very short wavelengths.

Other than parabolic antenna designs are also sometimes used in space telemetry. An omni-directional antenna which, ideally, radiates or receives energy equally in all directions is always included on a spacecraft as a safety factor to enable transmission to and from the vehicle regardless of its orientation in space. Evidently, this antenna has unity gain in all directions.

Clearly, stationary antennas can be made much larger than those that must be moved. Because space telemetry antennas must be accurately pointed in space, stationary antennas are not useful for this purpose. It is possible to get some effective direction change in stationary antennas by properly controlling the position of the source which radiates the antenna as well as the relative phases of the energy striking the various parts of the antenna surface. Such antennas are particularly useful in radio astronomy.

G. Analog Modulation

Before studying some of the more recent and sophisticated modulation techniques used in space telemetry communication systems, we shall present some of the more conventional techniques of wireless long distance communications. Typically, a signal of the form

$$A \cos (\omega t + \theta)$$

is generated at the transmitter. If the frequency $f = \omega/2\pi$ is sufficiently high, this signal can be applied to an antenna and will cause an electromagnetic wave to be emitted into space. An electromagnetic signal

$$B \cos [\omega(x - ct) + \theta]$$

will then be present at a receiving antenna, where $k = B/A$ represents the attenuation due to the medium and the distance through which the signal has traveled, and c the delay representing the time needed for the signal to travel from the transmitter to the receiver. In particular k does not vary with time, except perhaps for a slow steady change due to a change in the distance between the transmitter and the receiver.

If A , ω , and θ are kept constant at the transmitter, virtually no information can be transmitted. The receiver is able to determine that the transmitter must exist, but essentially nothing else. If, on the other hand, any one, or a combination of these parameters is varied in accordance with some rule known at both the transmitter and receiver, information can be transmitted. Commonly, there is some time function $m(t)$ which represents, for example, a temperature or pressure reading on a space vehicle, or a sound or light intensity as in commercial broadcasting. Thus, for example, if A is made to vary proportionally with $m(t)$, i. e.

$$A(t) = a m(t)$$

the resulting "amplitude modulation" signal is capable of conveying information. Similarly, if

$$\theta(t) = b m(t)$$

or if

$$\omega(t) = c m(t)$$

the signal is said to be "phase modulated" and "frequency modulated" respectively. These three types of modulation will now be examined.

H. Amplitude Modulation

The generation of an amplitude modulated signal is relatively straightforward. The signal $m(t)$ is converted to a voltage intensity in accordance with its amplitude (for example: a sound wave is passed through a microphone). This voltage is amplified and multiplied by a signal

$$A \cos (\omega_c t + \theta)$$

In addition, for reasons which will become apparent shortly, some unmodulated signal is also added in. All of these procedures can be readily accomplished electronically. The resulting amplitude modulated signal

$$x(t) = A [1 + \lambda m(t)] \cos (\omega_c t + \theta)$$

is then transmitted. The parameter λ is referred to as the "modulation index".

The spectrum of the signal $x(t)$ is simply a frequency translation of the modulating signal $m(t)$. To illustrate this, let us suppose that $m(t)$ is the sinusoidal signal

$$m(t) = a \cos \omega_c t$$

Then

$$\begin{aligned} x(t) &= A [1 + \lambda a \cos \omega_c t] \cos (\omega_c t + \theta) \\ &= A \cos (\omega_c t + \theta) + \frac{A\lambda a}{2} \left\{ \cos [(\omega_c + \omega_c) t + \theta] \right. \\ &\quad \left. + \cos [(\omega_c - \omega_c) t + \theta] \right\} \end{aligned}$$

and a modulating signal at the frequency $F_c = \omega_c / 2\pi$ results in a modulated signal leaving power at the carrier frequency F_0 at the "sum" frequency $F_c + F_c$ and at the difference frequency $F_c - F_c$. The power spectrum of the modulating signal

$$S_m(f) = \left| \mathcal{F} \{ m(t) \} \right|^2$$

is represented in Fig. 8.1(a) and that of the modulated signal

$$S_X(F) = \left| \mathcal{F} \{ x(x) \} \right|^2$$

in Fig. 8.1(b).

If F_m were the maximum frequency and F_{m1} the minimum frequency of the modulating signal (as indicated for example, in Fig. 8.2(a)) the frequency range of the modulated signal would extend from $(F_c - F_m)$ to $(F_c - F_{m1})$ and from $(F_c + F_m)$ to $(F_c + F_{m1})$, as in Fig. 8.2(b). The bandwidth due to amplitude modulation has consequently been doubled.

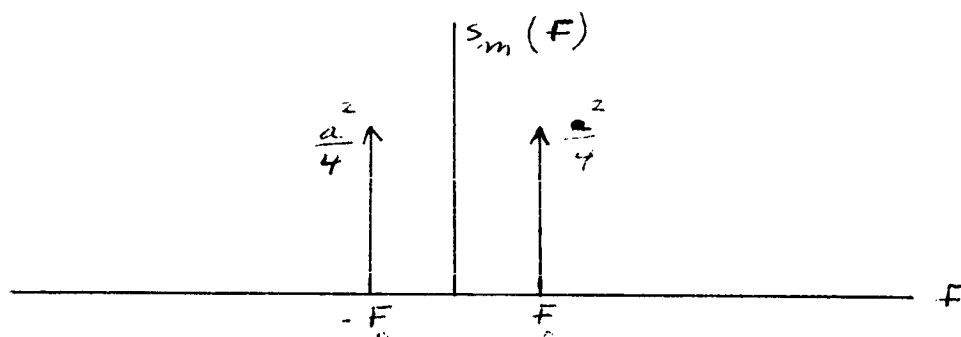


Fig. 8.1a Power spectrum of a sinusoidal modulation signal.

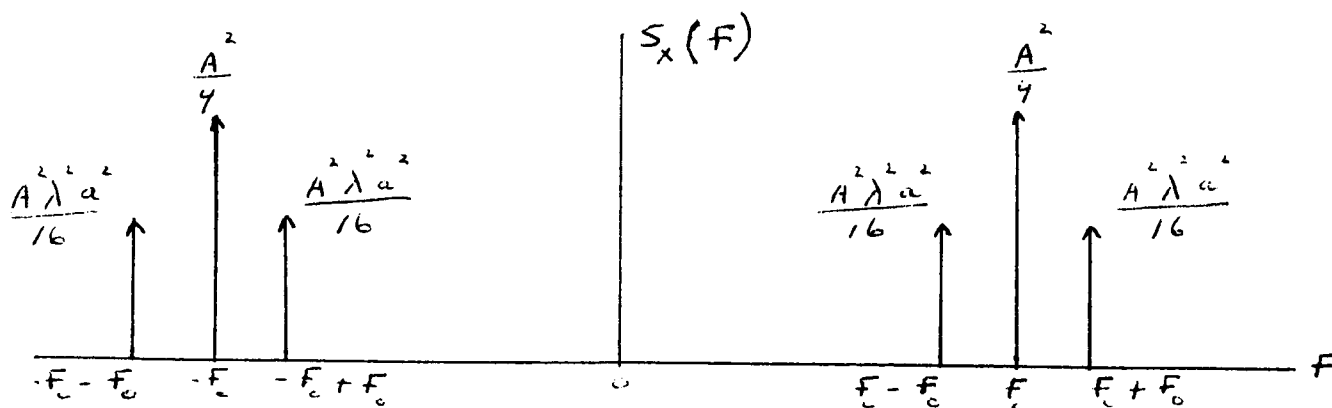


Fig. 8.1b Power spectrum of a carrier at frequency F_c amplitude modulated by a sinusoid of frequency F_0 .

Again, if the modulating signal were a sinusoid, the average power would be

$$P_{av} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x^2(t) dt = \int_{-\infty}^{\infty} S_x(f) df$$

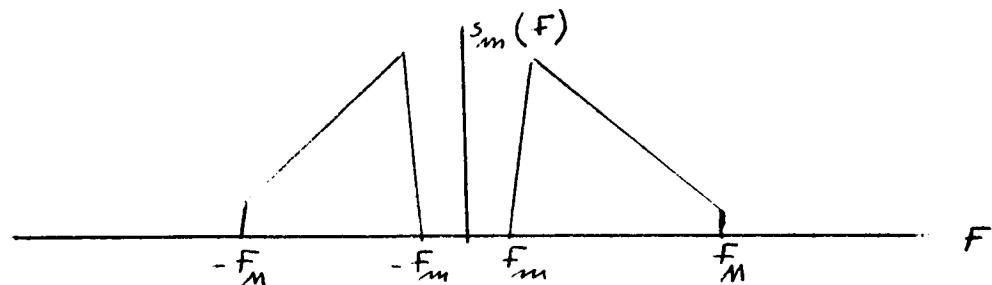
$$= \frac{A^2}{2} + \frac{A^2 \lambda^2 a^2}{4}$$

The first term in the equation for P_{av} represents the carrier power while the second represents $\frac{1}{2} A^2 \lambda^2$ times the power in the modulating signal. For a general modulating signal $m(t)$ the expression for the average transmitted power is

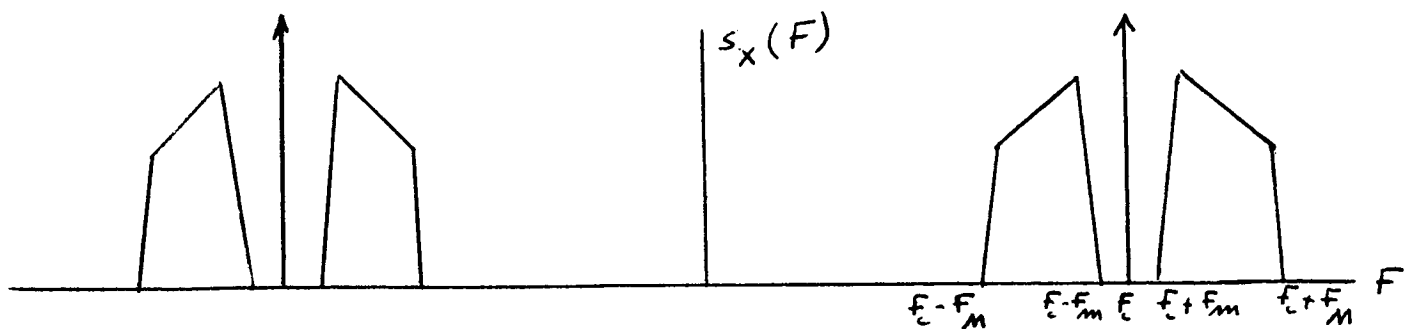
$$P_T = \frac{A^2}{2} + \frac{1}{2} A^2 \lambda^2 P_m$$

where P_m is the average power in the modulating signal. The percentage of the total power in the modulation is

$$\% \text{ Power in Modulation} = \frac{\frac{1}{2} A^2 \lambda^2 P_m}{\frac{1}{2} A^2 + \frac{1}{2} A^2 \lambda^2 P_m} = \frac{\lambda^2 P_m}{1 + \lambda^2 P_m}$$



(a) Power spectrum of a modulating signal.



(b) Modulated signal spectrum.

Fig. 8.2 Spectrum translation due to amplitude modulation.

In order to obtain useful information from the signal $x(t)$ at the receiver, it is necessary to "demodulate" it to obtain the desired signal $m(t)$. An AM signal may be demodulated in a number of ways. The most common technique involves the use of a non-linear element called a half-wave rectifier followed by a filter. The ideal half-wave rectifier may be regarded as a device whose output $y(t)$ is related to the input $x(t)$ as follows:

$$y(t) = \begin{cases} x(t) & \text{if } x(t) \geq 0 \\ 0 & \text{if } x(t) < 0 \end{cases}$$

A typical modulated waveform is shown in Fig. 8.3. It is seen that so long as $m(t) > -1$, the signal $x(t)$ is positive for $(2k - \frac{1}{2})T < t < (2k + \frac{1}{2})T$ and negative otherwise, where $T = 2\pi/\omega_c$.

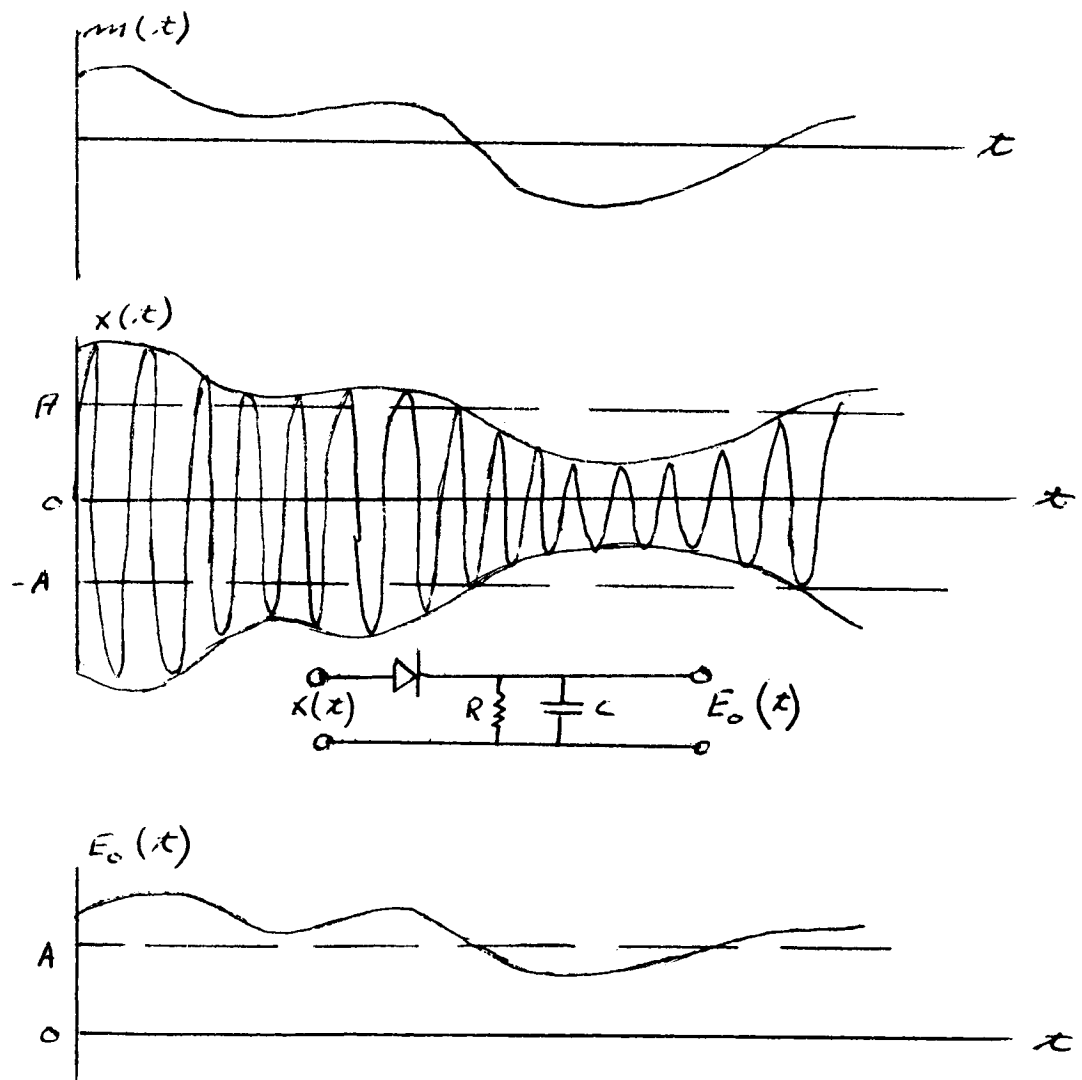


Fig. 8.3 Amplitude modulation and envelope detection.

Consequently, the half-wave rectifier has the same effect on the received waveform as if it were multiplied by the square wave in Fig. 8.4. As previously established, the Fourier series expansion of this square wave $S(t)$ is

$$S(t) = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{\sin \frac{n\pi}{2}}{\frac{n\pi}{2}} \cos n\omega_c t$$

Then

$$\begin{aligned} x(t) S(t) &= A (1 + \lambda m(t)) \cos \omega_c t \left[\frac{1}{2} + \sum_{n=1}^{\infty} \left(\frac{\sin \frac{n\pi}{2}}{\frac{n\pi}{2}} \right) \cos n\omega_c t \right] \\ &= A [1 + \lambda m(t)] \left\{ \frac{1}{2} \cos \omega_c t + \frac{1}{2} \sum_{n=1}^{\infty} \left(\frac{\sin \frac{n\pi}{2}}{\frac{n\pi}{2}} \right) [\cos (n-1)\omega_c t + \cos (n+1)\omega_c t] \right\} \end{aligned}$$

which is just a sum of amplitude modulated signals. The product $x(t) S(t)$, therefore contains terms centered about the frequencies $0, F_c, 2F_c, 3F_c, \dots$ as indicated in Fig. 8.5.

This signal is then filtered to remove the high frequency components, resulting in an output waveform $E_o(t)$ which ideally would be

$$E_o(t) = A [1 + \lambda m(t)]$$

as indicated in Fig. 8.3. That is, all but the desired term $A [1 + \lambda m(t)]$ can be eliminated by filtering, so long as

$$F_c - F_M > F_M$$

or equivalently if

$$F_M < \frac{F_c}{2}$$

where F_M is the highest frequency in the modulating signal.

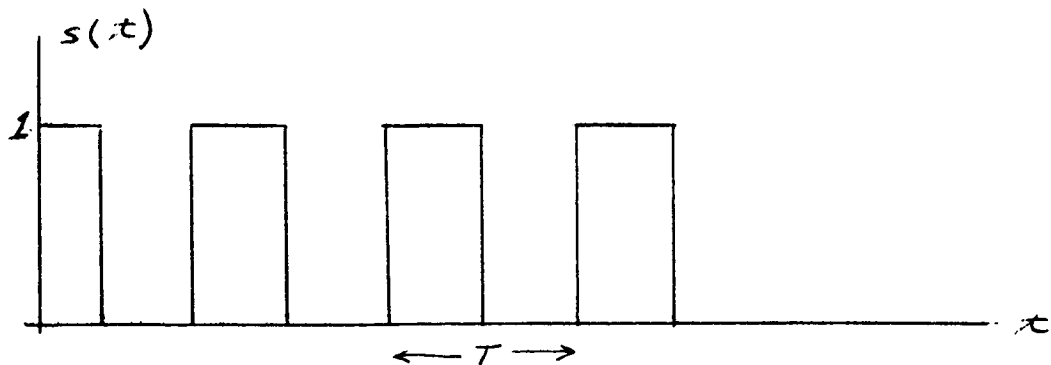


Fig. 8.4 A square wave of period $T = 1/F_c$.

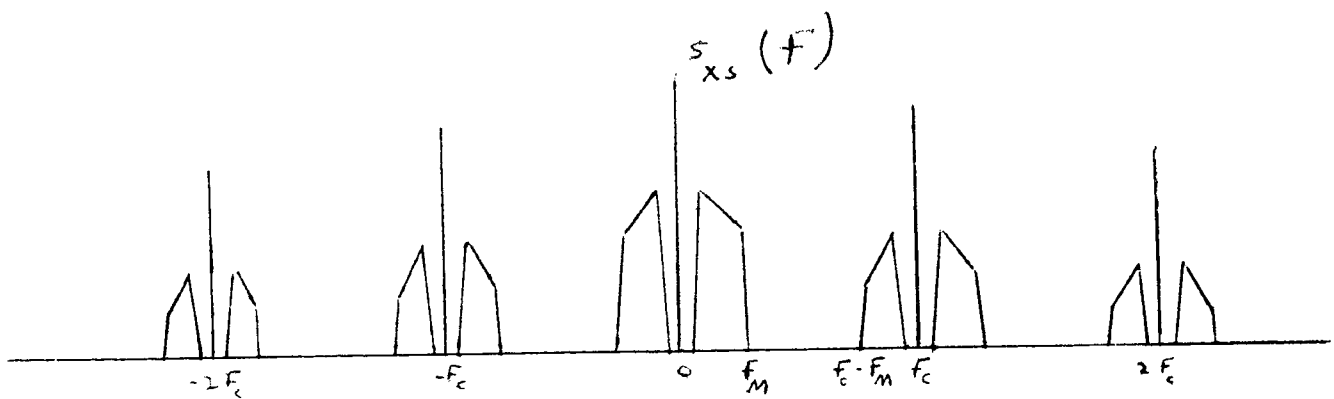


Fig. 8.5 Spectral density of the product $x(t) s(t)$.

Envelope detection of an amplitude modulated waveform is the simplest technique that may be employed. The main disadvantage with this method is that it is inefficient since the received waveform is shorted by the detecting diode whenever the wave has a negative voltage. One method of improvement is by the use of a full wave rectifier as the detector.

The common AM receiver used commercially today is the superheterodyne radio receiver, a typical diagram of which is shown in Fig. 8.6. The incoming signal passes through a tuned radio-frequency (r-f) amplifier which can be tuned variably over the radio band 550 to 1600 KC. This signal is then mixed with a locally generated signal. The sum and difference frequencies are generated as above and contain a term centered at 455KC.

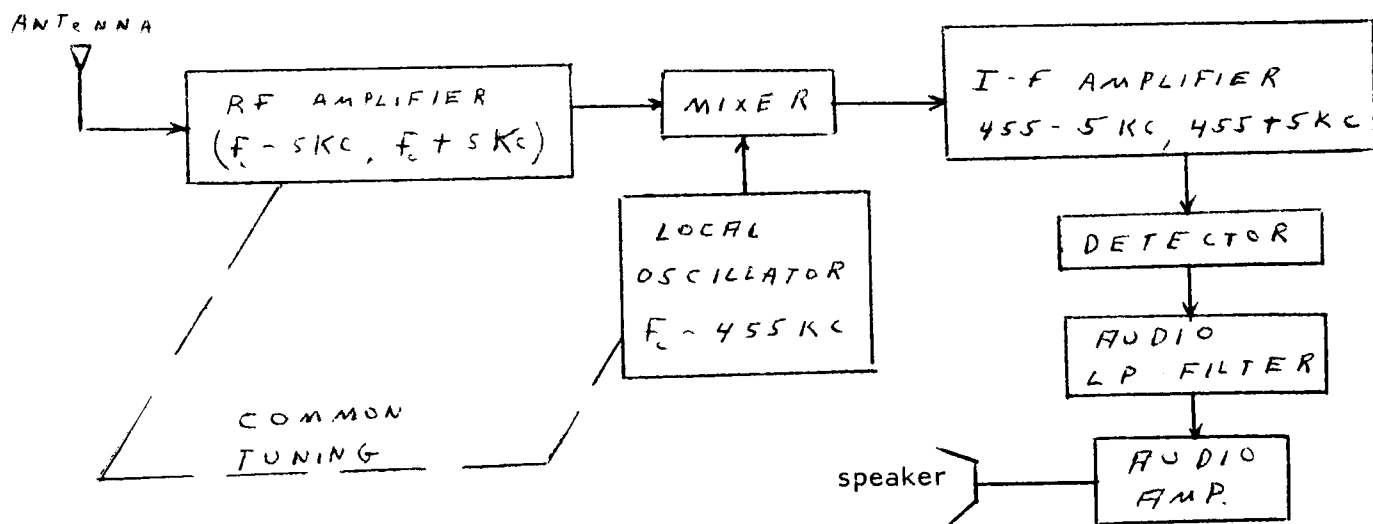


Fig. 8.6 Superheterodyne AM receiver.

The local oscillator and r-F amplifier are tuned together so that there is always a difference frequency of 455KC/sec between them. The mixer acts as a frequency converter, shifting the incoming signal down to the fixed intermediate frequency of 455KC/sec. Several stages of amplification are ordinarily used, with double-tuned circuits providing the coupling between stages. The intermediate frequency (i-F) signal is then detected as described above, amplified further in the audio-frequency (a-F) amplifiers and applied to the loudspeaker. The superheterodyning operation refers to the use of a frequency converter and fixed tuned i-F amplifiers before detection.

1. Double and Single Sideband Modulation

While the method of AM communications described in the previous section is quite satisfactory for commercial use, it has some limitations in those situations in which the available power is limited. First it will be recalled, the demodulation scheme outlined requires that

$$\lambda m(x) > -1$$

To appreciate the significance of this limitation, suppose

$$m(x) = \cos 2\pi F_0 t$$

then

$$P_m = \frac{1}{2}$$

and if the modulation index, λ , is set at its maximum value, namely 1, we see that the percentage of power in the modulation is at most 33 1/3%.

To overcome this difficulty, consider the following demodulation scheme: A narrow-band filter is centered about F_c and the output is used to estimate the frequency and phase of the carrier.

Suppose the carrier is of the form $A \cos \omega_c t$ and the estimate

$$c(t) = C \cos(\omega_0 t + \theta)$$

is made where presumably θ is small. Then the signal can be demodulated by forming the product

$$\begin{aligned}
 x(t) c(t) &= A [1 + \lambda m(t)] (\cos \omega_c t) [C \cos(\omega_c t + \theta)] \\
 &= AC [1 + \lambda m(t)] [\cos \theta + \cos(2\omega_c t + \theta)]
 \end{aligned}$$

which, after filtering the $2\omega_c t$ terms out, yields the desired signal

$$AC [1 + \lambda m(t)] \cos \theta$$

Note that in this scheme no limitations have been placed on the maximum or minimum values that $\lambda m(t)$ may take on. The only requirement now is that there is enough power in the carrier to enable a good estimate of its frequency and phase. It is not apparent that this is superior to the previous AM system until we determine how much power must be included in the carrier for satisfactory results. In a later section we shall verify, however, that in a typical situation, less than 1% of the total power need be included in the carrier, thus allowing a substantial increase in performance over conventional AM. Above we indicated that in conventional AM, over 66% of the power is in the carrier. This technique of increasing the proportion of power in the modulation by suppressing the carrier is commonly referred to as "double-sideband, suppressed carrier modulation," DSB/SC.

An interesting modification of the double-sideband modulation system is obtained by a technique known as single-sideband modulation, SSB. We have noted that the power spectrum $S(f)$ of an amplitude modulated signal is symmetric about the carrier frequency. When $m(t)$ is a sinusoid of frequency f_m the amplitude modulated signal contains a term at the frequency $f_c + f_m$ and a term at the same amplitude at $f_c - f_m$. This same symmetry occurs regardless of the signal $m(t)$.

Conventional AM or DSB modulation involves the product $m(t) \cos \omega_c t$, which, as we have seen shifts the spectrum of the modulating signal shown in Fig. 8.2(a) to the position indicated by Fig. 8.2(b). Now suppose the signal $m(t) \cos \omega_c t$ is passed through an ideal band pass filter with the pass band

$$f_c + f_m < |f| < f_c + f_M$$

The output then has the spectrum illustrated in Fig. 9.1. If this filtered signal $X'(t)$ is transmitted and demodulated by forming the product

$$X'(t) \cos \omega_c t$$

and filtering out the high-frequency components, the resulting signal has the spectrum in Fig. 8.2(a). That is, this resulting signal is identical to the original modulating signal $m(t)$. But note that by filtering before transmission, as described, only half as much bandwidth is needed for SSB and for DSB modulation. As with DSB modulation, it is necessary to transmit some carrier power in order to demodulate a SSB signal. It can be shown that in the latter case however, the phase accuracy of the estimate need not be as great as before to ensure the same performance.

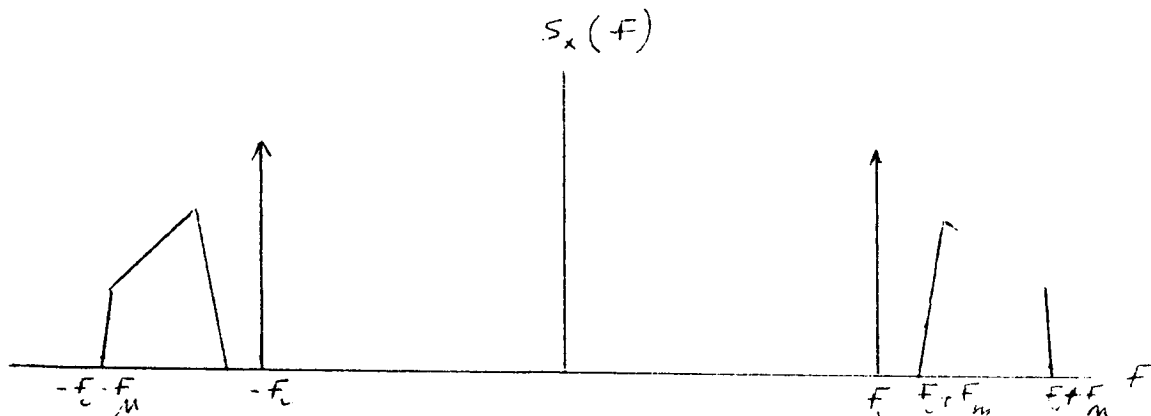


Fig. 9.1 Power spectrum for single side band modulation.

J. Noise Analysis of Amplitude Modulation Communications

The ultimate evaluation of any communication system rests in its behavior in the presence of noise. A convenient measure of this behavior is the output signal-power-to-noise-power ratio, often termed signal-to-noise ratio. In the case of DSB and SSB modulation combined with product demodulation, this ratio is readily determined. Consider first DSB modulation. (When we refer to DSB modulation we intend double-sideband suppressed-carrier modulation. The DSB/SC designation is somewhat redundant since we denote non-suppressed carrier modulation by "conventional AM.") This

received signal may be written

$$x(t) = A m(t) \cos \omega_c t + a \cos \omega_c t$$

The total power in the modulation is $\frac{1}{2} A^2 P_m$ where P_m represents the power in the modulating signal. The signal $x(t)$ is demodulated by forming the product $x(t) \cos \omega_c t$ and passing it through a low-pass filter with the cut-off frequency $B = F_m$. The output due to the signal is therefore

$$\left\{ [A m(t) \cos \omega_c t] [\cos \omega_c t] \right\}_{\substack{\text{Low} \\ \text{pass = L.F.}}} = \frac{A}{2} m(t)$$

where the subscript l.f. designates the low frequency components only. The output signal power is consequently $\frac{1}{4} A^2 P_m$, the power in the modulation.

The output noise signal is

$$n_1(t) = n(t) \cos \omega_c t$$

which simply represents a frequency translation of the noise $n(t)$. Under the assumption that the input is white, it remains white after the product is formed and since the average power in the sinusoid is $\frac{1}{2}$, the power spectral density of $n_1(t)$ is one-half that of $n(t)$, i.e. $\frac{1}{2} N_0$. The output noise power is consequently $\frac{1}{2} N_0 F_m$ and the output signal-to-noise ratio DSB modulation is

$$\left(\frac{S}{N} \right)_{DSB} = \frac{\frac{1}{2} A^2 P_m}{\frac{1}{2} N_0 F_m} = \frac{P_T - \frac{1}{2} a^2}{\frac{1}{2} N_0 F_m}$$

where P_T is the total received signal power,

$$P_T = \frac{1}{2} A^2 P_m + \frac{1}{2} a^2$$

When SSB modulation is used, although half the signal spectrum is suppressed, the other half can represent twice the power as before, keeping the total radiated power the same. After forming the product

$$x_{SSB}(t) \cos \omega_c t$$

and filtering as before, it is evident that the situation is identical to that for DSB modulation. Hence

$$\left(\frac{S}{N} \right)_{SSB} = \frac{P_T - \frac{1}{2} b^2}{\frac{1}{2} N_0 F_m}$$

where b is the amplitude of the received unmodulated carrier. Since, generally, a^2 and b^2 can both be small compared to the modulated power

$$\left(\frac{S}{N}\right)_{DSB} \approx \frac{P_T}{\frac{1}{2} N_c F_M} \approx \left(\frac{S}{N}\right)_{SSB}$$

Note that in each case we are considering ideal systems in which the unmodulated carrier power is negligible. The received signal $x(t)$ is demodulated by forming the product $x(t) \cos \omega_c t$. The demodulation scheme using a half-wave rectifier, of course, will not achieve the performance indicated here, (although at high signal-to-noise ratios the two methods give essentially the same results for conventional AM). Because we are considering modulation from the viewpoint of space communications and not as applied to commercial radio and television, we are primarily concerned with how well a particular modulation scheme can be made to work, not how well it works using inexpensive, mass produced receivers. Thus we are only interested in the ideal system as analyzed above, which can, by the way, be approached quite closely in practice. This spares us the considerably greater difficulty of analyzing the signal-to-noise ratios resulting from the use of more common demodulators such as the half-wave rectifier.

K. Phase and Frequency Modulation

Communication systems are now studied in which the transmitted signal is

$$x(t) = A \cos \theta(t)$$

the "angle" $\theta(t)$ varying in accordance with the modulating signal. If we define the "instantaneous frequency" as the rate of change of the phase angle $\theta(t)$ then

$$\omega(t) = \frac{d\theta(t)}{dt}$$

Note that this corresponds to the intuitive notion of frequency when

$$\theta(t) = \omega t + \theta_0$$

When ω varies with time however, the intuitive definition of frequency is somewhat less clear.

A phase modulation system is one in which the phase angle $\theta(t)$ is allowed to vary with the modulating signal, $m(t)$:

$$\theta(t) = \omega_c t + \theta_0 + A_\theta m(t)$$

Frequency modulation on the other hand implies that the instantaneous frequency is made to vary with $m(t)$

$$\omega(t) = \omega_c + \Delta\omega m(t)$$

and since

$$\omega(t) = \frac{d\theta(t)}{dt}$$

we have

$$\theta(t) = \int_0^t \omega(\tau) d\tau = \omega_c t + \theta_0 + \Delta\omega \int_0^t m(\tau) d\tau$$

Then FM is essentially PM with the exception that the modulating signal in the latter is the derivative of that in the former. For this reason, the two types of modulation may be analyzed simultaneously as long as this difference is borne in mind.

As with any modulation scheme, one of the first considerations when an FM (or PM) signal is to be transmitted is that of its bandwidth occupancy. Consider the case in which the modulating signal is sinusoidal

$$m(t) = \cos \omega_m t$$

Then

$$\omega(t) = \omega_c + \Delta\omega \cos \omega_m t$$

and

$$\theta(t) = \omega_c t + \left(\frac{\Delta\omega}{\omega_m} \right) \sin \omega_m t$$

The quantity

$$\beta = \frac{\Delta\omega}{\omega_m}$$

is defined to be the "modulation index" and is of fundamental importance in FM systems. Hence,

$$x(t) = A \cos (\omega_c t + \beta \sin \omega_m t + \theta_0)$$

or equivalently we have

$$\begin{aligned} x(t) &= A \cos (\omega_c t + \theta_0) \cos (\beta \sin \omega_m t) \\ &\quad - A \sin (\omega_c t + \theta_0) \sin (\beta \sin \omega_m t) \end{aligned}$$

Suppose for the moment, that β is small, say less than $\pi/18$. Then $\beta \sin \omega_m t$ is always less than $\pi/18$ radians or about 10° , and to a good approximation

$$\left. \begin{aligned} \cos(\beta \sin \omega_m t) &\approx 1 \\ \sin(\beta \sin \omega_m t) &\approx \beta \sin \omega_m t \end{aligned} \right\} \beta < \frac{\pi}{18}$$

The frequency modulated signal then is approximately

$$\begin{aligned} x(t) &\approx A \cos(\omega_c t + \theta_0) \\ &\quad - \beta A \sin \omega_m t \sin(\omega_c t + \theta_0) \end{aligned}$$

which, except for a 90° phase shift in the modulated carrier, is exactly the same as if the amplitude modulated signal

$$x_{AM}(t) \approx A [1 + \beta \sin \omega_m t] \cos(\omega_c t + \theta_0)$$

were transmitted. In general, if the modulating signal is $m(t)$ the frequency modulated signal may be approximated, for small modulation indices, by the signal

$$\begin{aligned} x(t) &\approx A \sin(\omega_c t + \theta_0) \\ &\quad + A \Delta\omega \left[\int_0^t m(\tau) d\tau \right] \cos(\omega_c t + \theta_0) \end{aligned}$$

Note that β must be small for all frequencies of the modulating signal $m(t)$ for this approximation to be valid. Specifically if $m(t)$ has a frequency component $\sin \omega_m t$, then β must be small. When this situation holds, the modulation is referred to as "narrow band FM". This is because the bandwidth required is the same as that needed for conventional AM when the modulating signal is $m(t)$. In particular, if the highest frequency component of $m(t)$ is F_M , then the maximum frequency component of its integral is F_M and the narrow-band FM bandwidth is just $2F_M$, as it would be with conventional AM. When β is increased, however, it will be seen that the FM bandwidth can be considerably greater than that necessitated by AM.

Suppose now that $\beta = \Delta F / F_0$ is large. Then the frequency deviation ΔF is large compared to the modulating frequency F_0 . In this case, the frequency being transmitted varies from $F_c - \Delta F$ to $F_c + \Delta F$ and, importantly, it varies between these extremes at a rate which is slow compared to the distance

over which it varies. The transmitted signal could be approximated by a signal at the frequency $F_c + \Delta F$, followed some time later by a signal at the frequency $F_c + \Delta F - \delta F$, followed still later by $F_c + \Delta F - 2\delta F$ etc. The important feature of this signal is the frequency extremes through which it varies, not the relatively slow rate at which it changes frequencies. This suggests that the bandwidth required for FM, when β is large, is approximately

$$2\Delta F = 2\beta F_c \text{ instead of the value } 2F_c \text{ needed when } \beta \text{ is small.}$$

When β is on the order of 10 or greater, this estimate of the required bandwidth is quite accurate and it is reasonably applicable even when β is as small as four, as the following rigorous analysis will verify. Since, when β is large, the FM bandwidth is increased by a factor of β over that needed for AM, this modulation scheme is designated wide-band FM.

To rigorously analyze a sinusoidally modulated FM signal, we are basically interested in determining the frequency components of the frequency modulated carrier given by

$$x(t) = A \cos(\omega_c t + \phi_c) \cos(\beta \sin \omega_m t) - A \sin(\omega_c t + \phi_c) \sin(\beta \sin \omega_m t)$$

But we note that both $\cos(\beta \sin \omega_m t)$ and $\sin(\beta \sin \omega_m t)$ are periodic functions of $\omega_m t$. As such, each may be expanded in a Fourier series of period $2\pi/\omega_m$. Each series will contain terms in ω_m and all its harmonic frequencies. By writing

$$e^{i\beta \sin \omega_m t} = \cos(\beta \sin \omega_m t) + i \sin(\beta \sin \omega_m t)$$

the evaluation of these Fourier series may be carried out by determining only one set of coefficients. For this we write the representation

$$e^{i\beta \sin \omega_m t} = \sum_{n=-\infty}^{\infty} a_n e^{i n \omega_m t}$$

where

$$a_n = \frac{\omega_m}{2\pi} \int_0^{2\pi/\omega_m} e^{i\beta \sin \omega_m t} e^{-i n \omega_m t} dt$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\beta \sin y} e^{-i n y} dy$$

$$= J_n(\beta)$$

which is the n -th order Bessel function with the argument β . Bessel functions are already tabulated for different values of n and β and can therefore be assumed to be known. By equating real and imaginary terms, we obtain

$$\cos(\beta \sin \omega_c t) = J_0(\beta) + 2 J_2(\beta) \cos 2 \omega_c t + 2 J_4(\beta) \cos 4 \omega_c t + \dots$$

and

$$\sin(\beta \sin \omega_c t) = 2 J_1(\beta) \sin \omega_c t + 2 J_3(\beta) \sin 3 \omega_c t + \dots$$

Using these Fourier series expansions, and the utilizing the trigonometric sum and difference formulas, we obtain

$$\begin{aligned} x(t) = & J_0(\beta) \cos \omega_c t \\ & - J_1(\beta) [\cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t] \\ & + J_2(\beta) [\cos(\omega_c - 2\omega_m)t + \cos(\omega_c + 2\omega_m)t] \\ & - J_3(\beta) [\cos(\omega_c - 3\omega_m)t - \cos(\omega_c + 3\omega_m)t] + \dots \end{aligned}$$

We thus have a time function consisting of a carrier and an infinite number of sidebands, spaced at frequencies $\pm F_m$, $\pm 2 F_m$, $\pm 3 F_m$ etc. away from the carrier. This is in contrast to the AM case, when the carrier and only a single set of sidebands existed. The magnitudes of the carrier and sideband terms depend on β , the modulation index, this dependence being expressed by the appropriate Bessel function. Fig. 11.1 gives a plot of the several Bessel functions of the first kind.

In Fig. 11.2 is plotted an FM signal spectrum for a sinusoidal modulating signal for various modulation

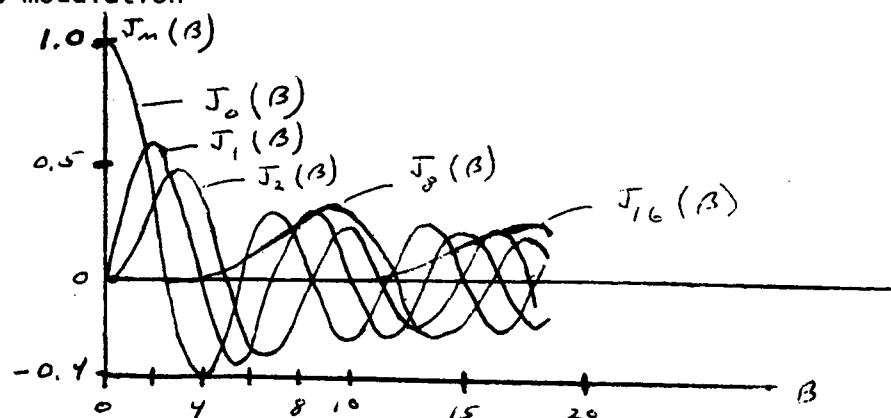


Fig. 11.1 Plots of Bessel functions of the first kind.

In Fig. 11.3 is plotted an FM signal for sinusoidal modulating signal for a fixed $\Delta F = \beta \omega_c$ for various β . These indicate the concentration of the spectral lines and allow excellent approximations to bandwidth. As noted, the bandwidth approaches $2\Delta F$ for large β . (The FCC has fixed the maximum value of ΔF at 75 KC for commercial FM broadcasting stations. Thus, if we take the maximum modulating to be 15 KC, as is typically used as the maximum audio frequency, then $\beta = 5$.)

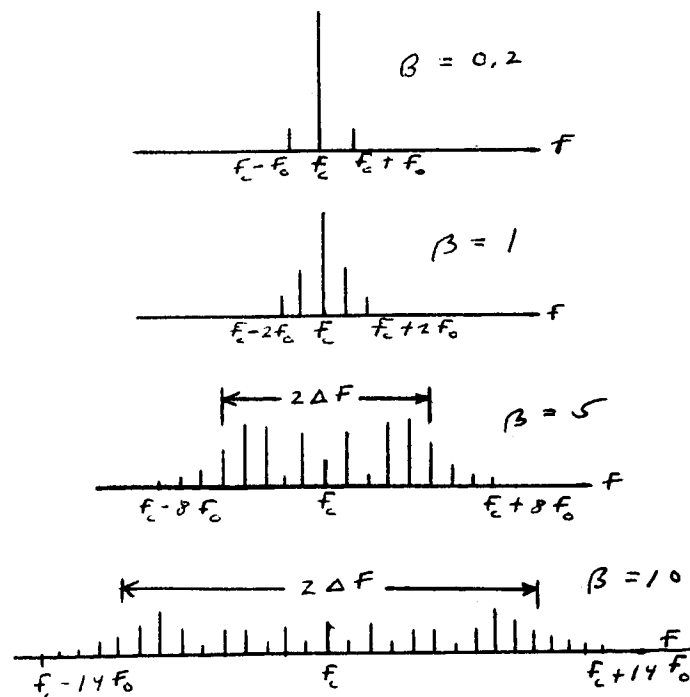


Fig. 11.2 Amplitude frequency spectrum - FM signal, f_m - fixed, amplitude varying

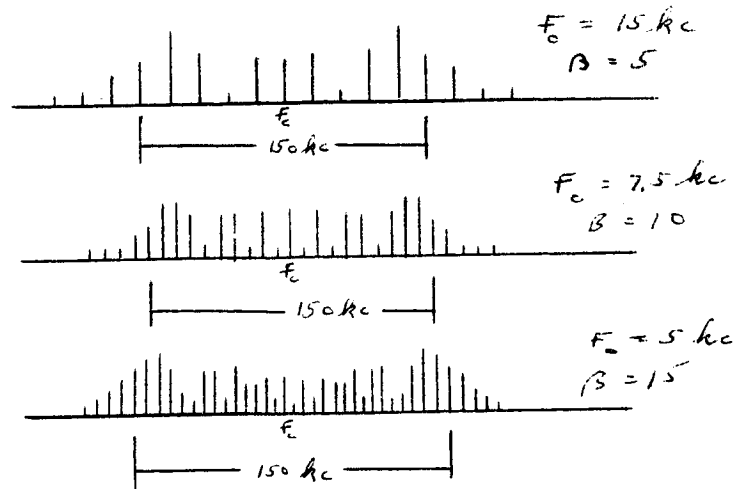


Fig. 11.3 Amplitude - Frequency spectrum - FM signal, AF fixed, f_o varying

Note that since $\beta = \frac{\Delta\omega}{\omega_c} = \frac{\Delta F}{F_c}$ in the case of frequency modulation, $B \approx 2\Delta F$ when β is large and is independent of the modulating frequency so long as the amplitude of the modulating signal does not vary with frequency causing an effective variation in ΔF . Thus if the average power in the modulating signal $m(t)$ is the same for all modulating frequencies, i.e. if the power spectrum of $m(t)$ is flat, then, on the average, the bandwidth occupancy of the FM signal will not vary with the frequency of the modulating signal. Since, as we shall show, the performance of FM is proportional to its bandwidth it is desirable to have maximum bandwidth occupancy as consistently as possible.

With a phase modulated signal, the analysis is identical except that now $\beta = \Delta\theta$ and the bandwidth is $F_c \Delta\theta$. Thus, if the amplitude of the modulating signal is independent of frequency, the bandwidth of a phase modulated signal increases with the modulating frequency, a generally less desirable situation. On the other hand, ordinary speech and music exhibit the property that the amplitude of a frequency component, beyond a certain frequency, tends to be inversely proportional to the frequency. In this case, $\Delta\theta \propto \frac{1}{F_c}$ and the bandwidth B of a PM signal remains constant, independent of frequency, whereas an FM bandwidth would decrease with increasing frequency. For this reason, commercial FM modulating signals are preceded by a "pre-emphasis" network

which increases the magnitude of the higher frequency components by an amount proportional to their frequency.

This is then counteracted by a "de-emphasis" network at the receiver which reverses the operation. Commercial FM therefore is strictly neither FM nor PM but a combination of both. Clearly, the distinction is irrelevant so far as the system is concerned, the only difference between the two being that of a preconditioning of the modulation signal.

Generally, then, an FM or PM modulation system is as illustrated in Fig. 11.4. A voltage controlled oscillator (VCO) is a sinusoidal oscillator, the output frequency of which is proportional to the input voltage, that is, if the input voltage is $m(t)$ volts, the output frequency is $f_c + \Delta f m(t)$ cps.

There are a number of ways of implementing the block diagram in Fig. 11.4. However, since we are interested primarily in the system aspects rather than in its particular implementation, suffice it to observe that voltage controlled oscillators can be designed to give the desired performance over a wide frequency range.

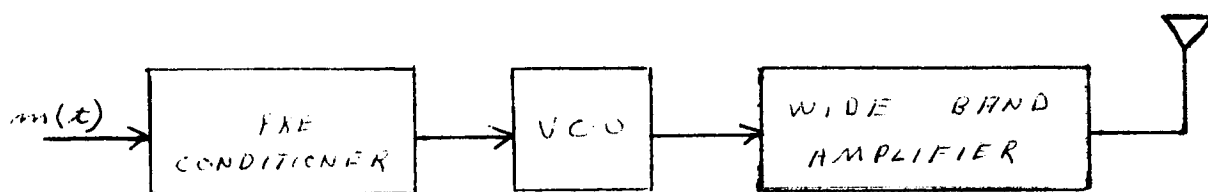


Fig. 11.4 An FM transmitter.

L. Demodulation of Phase and Frequency Modulation

There are a number of ways in which an FM signal may be demodulated. Any device which is capable of converting a frequency variation into an amplitude variation can serve as an FM demodulator. Such a device is called a frequency discriminator. Suppose, for example, that the FM signal is passed through a filter with the characteristics

$$H(f) = Kf$$

$$f_c - \Delta f < |f| < f_c + \Delta f$$

Clearly, the output amplitude is proportional to the input frequency desired and the FM signal is thereby demodulated. This, in fact, is a somewhat

simplified version of a commercial FM discriminator. In Fig. 11.5, there is a block diagram of a commercial FM receiver which will be described in more detail later.

Another FM demodulator can be designed from the following point of view: Suppose we have, at the receiver, a VCO which is identical to that at the transmitter. If we then make a preliminary estimate of the amplitude of the modulating signal and apply this to the VCO, the similarity between the output of the VCO and the received FM signal will provide us with a measure of the accuracy of this estimate. If we use the comparison itself to adjust the VCO the system can be made to "track" the modulating signal. An illustration of how this may be accomplished is given in Fig. 11.6. This device, called a "phase locked loop" consists of a multiplier, a filter $h(\tau)$, a VCO and a device which shifts the phase of the VCO output by -90° . To analyze its behavior suppose the signal $x(t)$ is

$$x(t) = A \cos(\omega_c t + \theta_1)$$

and suppose the VCO output is

$$\hat{x}_s(t) = \cos(\omega_c t + \theta_2)$$

where $\omega_c = \theta_1 - \theta_2$ represents a "small" tracking error. Then the product $x(t)\hat{x}_s(t) \cdot \epsilon$

where $\hat{x}_s(t)$ represents the shifted version of $\hat{x}(t)$ is formed, yielding

$$\begin{aligned} x(t) \hat{x}_s(t) &= [A \cos(\omega_c t + \theta_1)] [B \cos(\omega_c t + \theta_2 + \frac{\pi}{2})] \\ &= \frac{AB}{2} \left[\cos(\theta_1 - \theta_2 - \frac{\pi}{2}) + \cos(2\omega_c t + \theta_1 + \theta_2 + \frac{\pi}{2}) \right] \end{aligned}$$

The second term is a high-frequency correspondent and will be eliminated by the combined action of the VCO and the filter $h(\tau)$. The low-frequency component is

$$\cos(\theta_1 - \theta_2 - \frac{\pi}{2}) = \sin \theta_1 - \sin \theta_2 \approx \theta_1 - \theta_2$$

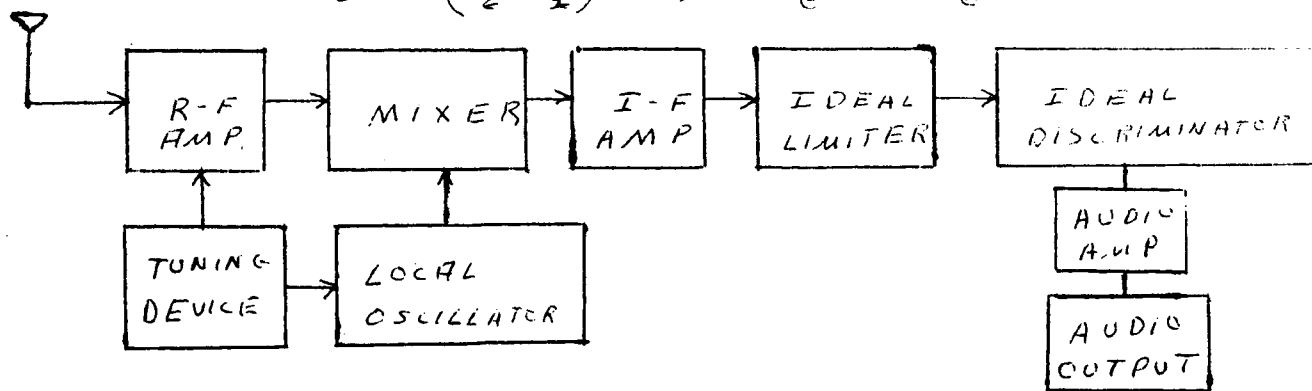


Fig. 11.5 Block diagram - commercial FM receiver.

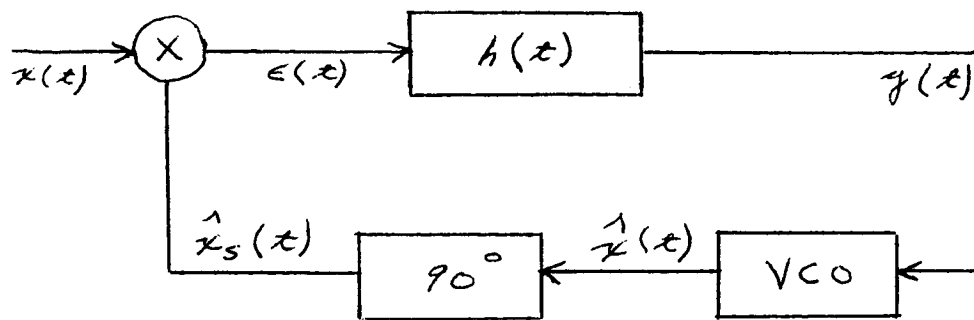


Fig. 11.6 A phase-locked loop.

The last approximation is based on the assumption that the phase error θ_e is small. Thus θ_e is the input to the VCO. Suppose θ_e is positive. Then the VCO frequency is increased to something slightly greater than ω_c thereby decreasing the difference between θ_1 and θ_2 and hence decreasing θ_e . Similarly, if θ_e is negative, the VCO frequency is decreased, again decreasing the absolute value of the difference between θ_1 and θ_2 . The loop therefore acts to reduce the phase error to zero.

Now suppose θ_1 varies with time, $\theta_1 = \theta_1(t)$. The loop will again act in such a way as to keep the phase error nearly zero. Then

$$\theta_2(t) \approx \theta_1(t)$$

The difference between the VCO center frequency ω_c and its actual frequency is proportional to its voltage input. Since the instantaneous frequency of the VCO output is

$$\frac{d}{dt} [\omega_c t + \theta_2(t)] = \omega_c + \frac{d\theta_2(t)}{dt}$$

the input to the VCO must have amplitude

$$k \frac{d\theta_2(t)}{dt} \approx k \frac{d\theta_1(t)}{dt}$$

where k is a constant of proportionality. Consequently, if the input to the loop is a frequency modulated signal

$$\theta_1(t) = \Delta\omega \int^t m(\tau) d\tau + \theta_0$$

then the input to the VCO is just

$$y(t) = k \frac{d\theta_1(t)}{dt} = k \Delta\omega m(t)$$

M. Noise Analysis of FM and PM

In order to determine the effect of noise at the input, let $x(t) = m(t)$ be white noise. Then

$$m(t) \hat{x}_s(t) = m(t) A \cos \left(\omega_c t + \theta_1(t) + \frac{\pi}{2} \right) = m_1(t)$$

is also white noise, as was previously observed, and with the same power spectral density, N_0 , as $m(t)$.

Now consider the situation in which

$$x(t) = A \cos \left(\omega_c t + \theta_1(t) \right) + m(t)$$

and

$$\hat{x}_s(t) = \cos \left(\omega_c t + \theta_2(t) + \frac{\pi}{2} \right)$$

where again, it is assumed $\theta_1(t) - \theta_2(t)$ is small. The low frequency term of the product $x(t) \hat{x}_s(t)$ is given by

$$\begin{aligned} x(t) \hat{x}_s(t) \Big|_{L.F.} &= A \sin \left(\theta_1(t) - \theta_2(t) \right) + m_1(t) \\ &\approx A \left[\theta_1(t) - \theta_2(t) \right] + m_1(t) \end{aligned}$$

Since $\theta_2(t)$ is to be adjusted by the action of the loop to keep the error signal and hence the input to $h(t)$ small, it follows that

$$\theta_2(t) \approx \theta_1(t) + \frac{m_1(t)}{A}$$

and hence

$$y(t) = k \frac{d\theta_2(t)}{dt} \approx k \frac{d\theta_1(t)}{dt} + \frac{k}{A} \frac{dm_1(t)}{dt}$$

Since the desired output is $k \Delta \omega m(t)$ the term $\frac{k}{A} \frac{dm_1(t)}{dt}$ represents output noise.

To determine the effect on the noise of taking its derivative, suppose that it consists of a single frequency component

$$m(t) = a_m \cos (\omega_m t + \phi_m)$$

Then the derivative of the noise consists of the component

$$-a_m \omega_m \sin (\omega_m t + \phi_m)$$

and hence has a magnitude equal to ω_m times the magnitude of the original noise.

In general, if the power spectral density of the white noise is N_0 cps for all f , the power spectral density of its derivative is just $(2\pi f)^2 N_0$.

The signal-to-noise ratio at the output of the FM demodulator is determined as follows: The power in the received signal $k \Delta \omega m(t)$ is, of course

$$S_m = k^2 (\Delta \omega)^2 P_F$$

where P_F is the power in the modulating signal. For the noise

$$S_N = \left(\frac{k}{A}\right)^2 N_0 \int_0^B (2\pi F)^2 dF = 4\pi^2 \left(\frac{k}{A}\right)^2 N_0 \frac{B^3}{3}$$

where B is the bandwidth of the output signal. Clearly $B = F_M$ the maximum frequency component of the modulating signal, since no higher frequencies are of interest. If the loop itself did not eliminate all frequencies greater than F_M cps, it could be followed by a low-pass filter which did. So the output signal-to-noise ratio is

$$\begin{aligned} \left(\frac{S}{N}\right)_{FM} &= \frac{3A^2 (\Delta \omega)^2 P_F}{(2\pi)^2 F_M^3 N_0} = 3 \left(\frac{\Delta \omega}{\omega_M}\right)^2 \frac{A^2 P_F}{N_0 F_M} \\ &= 3 (\beta_M)^2 \left(\frac{S}{N}\right)_{SSB} \end{aligned}$$

Recalling that, for large modulation indices, β_M may be interpreted as the ratio of the bandwidth needed with FM to that necessary for conventional AM or DSB transmission, we see that the signal-to-noise ratio improvement in FM is proportional to the square of this bandwidth multiplication factor β_M . Consequently, FM provides a means to obtain improved performance by increasing the signal bandwidth. Since increasing the FM bandwidth by β achieves the same results as increasing the signal power by β^2 , FM may also be regarded as a method of exchanging power for bandwidth to keep the same performance.

The analysis of the signal-to-noise performances of PM follows along the same lines as that for FM, except that rather than the signal $y(t)$ of Fig. 11.6, we are interested in its integral. That is, since

$$\theta_2(t) \approx \theta_1(t) + \frac{m_1(t)}{A}$$

where, in this case, $\theta_1(t) = \Delta \theta m(t)$, it follows that $\theta_2(t)$ is the quantity of interest, not its derivative $y(t)$. Thus the input to the VCO must be passed through an integrator in order to yield the desired output. The output signal power is

$$(\Delta \theta)^2 P_F$$

while the noise power is

$$\frac{1}{A^2} N_0 F_M$$

resulting in a signal-to-noise ratio

$$\left(\frac{S}{N}\right)_{PM} = (\Delta\theta)^2 \frac{A^2 P_F}{N_o F_M} = (\Delta\theta)^2 \left(\frac{S}{N}\right)_{SSB}$$

In the above discussion of phase-locked-loop demodulation of FM and PM, we have made some assumptions which should be emphasized. In particular, it was assumed that conditions were such that the VCO phase output was sufficiently close to the input phase that the approximation

$$\sin(\theta_1(t) - \theta_2(t)) \approx \theta_1(t) - \theta_2(t)$$

was valid. However, the loop dynamics require that

$$\theta_2(t) \approx \theta_1(t) + \frac{n_1(t)}{A}$$

Clearly, if the term $\frac{n_1(t)}{A}$ represents a phase angle of, say, more than 10° , then this approximation becomes unacceptable. But since the power of the normalized noise $\frac{n_1(t)}{A}$ within the bandwidth F_M of the signal is $\frac{N_o F_M}{A^2}$, it follows that $\frac{1}{A} n_1(t)$ will be small compared to 10° only so long as the term $\frac{N_o F_M}{A^2}$ is sufficiently small. As described, this analysis under these ideal assumptions is a linear analysis and as such is an approximation to the inherently non-linear phase locked loop. In general, it is necessary to require that

$$\frac{A^2}{N_o F_M} \geq 36$$

in order that all the assumptions made are reasonable. While the demodulated signal may be meaningful for smaller value of $A^2/N_o F_M$ than 36, the performance rapidly deteriorates as this ratio is further decreased.

Also, we have not specified the form of the filter $h(t)$. Clearly, it is to be chosen if possible so that $\theta_2(t) \approx \theta_1(t)$ regardless of the variation of $\theta_1(t)$ even in the presence of the noise $n_1(t)$.

Techniques are available for mathematically specifying the optimal filter when the signal and noise spectral densities are known. Nevertheless, if the normalized noise power $N_o F_M/A^2$ is large, the difficulties mentioned above remain, regardless of the filter $h(t)$.

This threshold effect when the noise becomes sufficiently large, is characteristic of any FM demodulating scheme. Heuristically, it can be said that, since the information conveyed in an FM signal is by the instantaneous frequency, a

measure of the effect of the noise exists in the comparison of the position of the zero crossings before and after the addition of the noise. It can be observed that, as the noise increases, some zero crossings will be added by the additive noise, while others will be eliminated entirely. When the noise reaches a level at which this phenomena becomes relatively common, the demodulated signal rapidly deteriorates.

An effective device which is used to reduce the noise is the limiter. Since the information in the FM signal is in the instantaneous frequency, and hence in the rate of zero crossings, and not in the signal amplitude, the signal can be made to approach its original form by limiting its amplitude. This, of course is accomplished without altering the position of the zero crossings. Hence, the ideal limiter is inducted in commercial FM receivers, as indicated in Fig. 11.5.

N. Other Applications of Phase-Locked-Loops

We now mention some applications of phase-locked loops other than FM and PM modulation.

In the case of DSB and SSB modulation, it was suggested that the suppressed carrier be tracked by a phase-locked loop in order to acquire a reasonably accurate estimate of it which is necessary for product demodulation. The analysis of the phase-locked loop in this situation is identical to that presented in the previous section with two exceptions: 1) the phase of the received signal $\theta_r(t)$ is constant except for a small variation caused by instabilities in the transmitter oscillator, movement of the transmitter relative to the receiver, and perhaps by random fluctuations caused by transmission medium. It is not caused to vary deliberately, and hence the bandwidth of $\theta_r(t)$ is very much less here than in the case of FM demodulation. 2) the desired signal is not $y(t)$ but rather $\hat{x}(t)$ since it is the carrier itself, not just its phase which is to be estimated. The phase error of the estimate, $\hat{x}(t)$, it is seen, is just $n_e(t)/a$, where a is the carrier amplitude and represents an effective phase error power $N_e B_L / a^2$ where B_L is the loop bandwidth. Thus, since the loop bandwidth B_L can be made very narrow, the phase error can be reasonably small, even for quite small values of a . Since the phase error power $N_e B_L / a^2$ is the expected value of the square of the phase error, the square root of this quantity gives an estimate of the magnitude of the phase error which will be encountered. By requiring $\sqrt{N_e B_L / a^2}$ to be less than 1/6 radian, for

example, one can be reasonably sure that the phase error remains within tolerable limits. It will be recalled that

$$\left(\frac{S}{N}\right)_{SSB} = \frac{P_T}{N_c F_M}$$

where P_T is the total power in the received signal, N_c , the noise spectral density, and F_M the signal bandwidth. This was true under the assumption that the ratio of the power in the carrier to that in the modulation was negligibly small. Suppose, as an example, that it is required that the output signal-to-noise ratio $\left(\frac{S}{N}\right)_{SSB}$ must be at least four, i.e. the signal power must be at least four times as great as the noise power, a somewhat marginal condition. Then

$$\frac{N_c B_L}{a^2} \left(\frac{S}{N}\right)_{SSB} = 4 \left(\frac{1}{36}\right)$$

or

$$\frac{N_c B_L}{a^2} \frac{P_T}{N_c F_M} = \frac{1}{9}$$

and

$$\frac{P_T}{a^2} = \frac{1}{9} \frac{F_M}{B_L}$$

Typically $F_M = 6000 \text{ cps}$ while the effective loop bandwidth of the carrier tracking loop can be made 1.0 cps or less. Thus

$$\frac{P_T}{a^2} = \frac{2}{3} \cdot 10^3$$

and indeed, the required carrier power is negligible.

These same comments are equally applicable to any filter of bandwidth B_L . The advantage of the phase-locked loop over ordinary filters is that it is able to track the signal, thereby requiring a much smaller bandwidth. Suppose, for example, that in the course of a transmission the carrier frequency could vary by as much as 20 cps, actually a reasonably small variation when all possible causes, such as doppler shifts and transmitter instabilities are taken into consideration. A conventional filter would need a bandwidth of at least 20 cps in order not to lose the signal entirely. If the rate at which the frequency varied were small, however, the phase-locked loop could track it typically with a bandwidth of 1 cps or less.

0. Pulse Modulation

We now consider digital modulation techniques as opposed to the analog modulation systems discussed previously. Another term often used is pulse modulation. Such systems are often used in telemetry communication systems which is defined as a system which measures pressure, temperature, radiation, and possibly other physical quantities, and transmits the information to a distant receiver. To start, we restate the sampling theorem discussed previously.

The sampling theorem states that we need only concern ourselves with periodic samples of a strictly bandwidth limited time function. Only the sequence of numerical values $x(nT)$ need be transmitted. The complete function may be reconstructed at the receiver by generating a series of delta function of area $x(nT)$ and passing them through a lowpass filter.

There are a number of advantages associated with sampled data telemetry systems. We shall see that there are a number of elegant techniques that have been devised for transmitting sampled data, methods which simultaneously are relatively easily implemented and achieve large signal-to-noise ratio. This is usually accomplished at the expense of additional bandwidth.

Also, it is often considerably easier and more efficient to handle sampled data than continuous data. Typically, a spacecraft may contain 100 to 1000 data sources. Some method must be employed to keep the information from each source separate. One method for doing this is called frequency multiplexing, which consists of forming the products

$$x_i(t) \cos \omega_i t$$

for each data signal $x_i(t)$, $i = 1, \dots, N$. The frequencies $f_i = \omega_i / 2\pi$ must be chosen such that the spectra of each signal does not overlap, as shown in Fig. 15.1. The waveform

$$x(t) = \sum_{i=1}^N x_i(t) \cos \omega_i(t)$$

then has the composite spectrum in Fig. 15.1 and $x(t)$ may be treated as a single source with a bandwidth

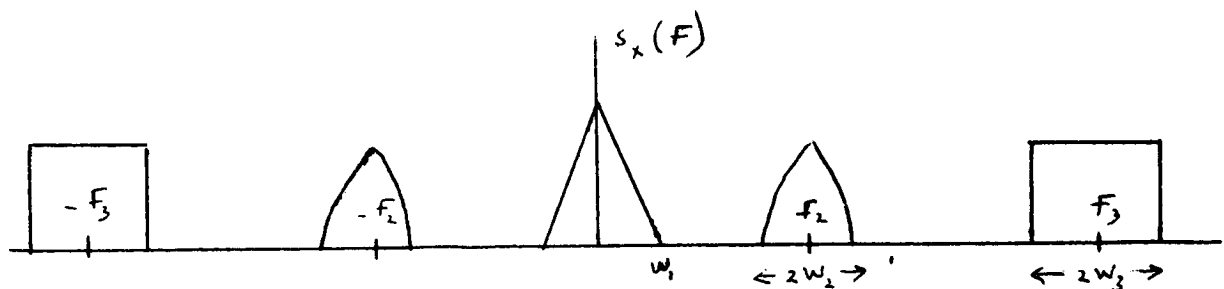


Fig. 15.1 Composite spectrum of frequency multiplexing.

$$W = W_1 + 2 \sum_{i=2}^N W_i \quad (15.1)$$

where W_i is the bandwidth of the process $x_i(t)$. Since the individual spectra do not overlap, the different signals $x_i(t)$ can be reconstructed at the receiver by proper filtering. Unfortunately, this method demands that each source be followed by a device for forming the product $x_i(t) \cos \omega_i t$, a procedure which is quite inefficient, particularly on board a spacecraft. (It is adequate and is used in commercial FM stereo multiples transmission, for which case $N = 2$).

The alternative is to sample each of the signals $x_i(t)$, represent the samples as pulses of duration T/N where T is the sampling rate. We here assume that all signals have the same bandwidth so that $T = 1/2W$ is the same for all $x_i(t)$. Finally time multiplex the samples as indicated in Fig. 15.2. The pulse labeled i corresponds to a sample of the process $x_i(t)$. If the bandwidth of the pulses are not the same, the sampling rates must be different, or all rates must be equal to that required by the signal with the largest bandwidth. Different sampling rates can readily be accommodated as long as they are integrally related. That is, suppose for example, $x_1(t)$ has a bandwidth which is twice as great as $x_2(t)$ and the two are to be time multiplexed. Since $x_1(t)$ must be sampled twice as often as $x_2(t)$ they can be multiplexed as in Fig. 15.3. Thus, in time T , two samples of $x_1(t)$ are transmitted while only one sample of $x_2(t)$ is transmitted, and both are sampled at the correct rate.

A time multiplexing system involves only the problem of commutation or the interspacing of samples from the various sources at the proper rate. No power consuming auxiliary equipment is required other than a moderate number of switching devices.

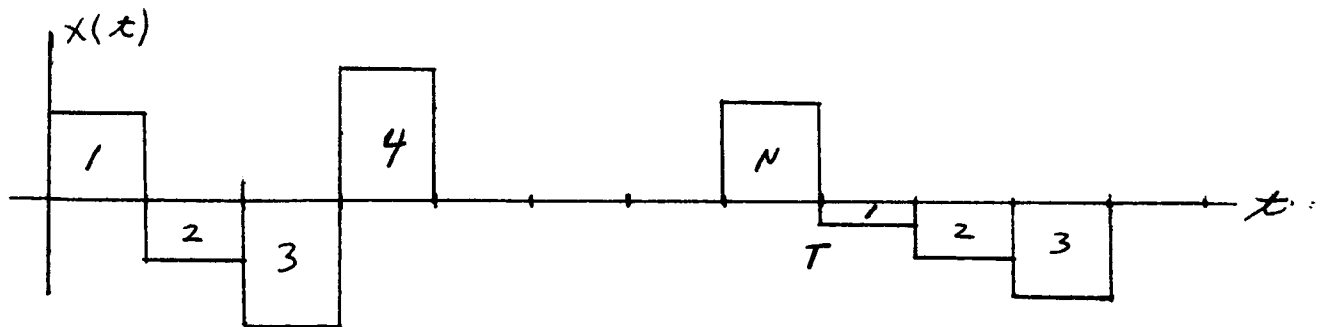


Fig. 15.2 Time multiplexing.

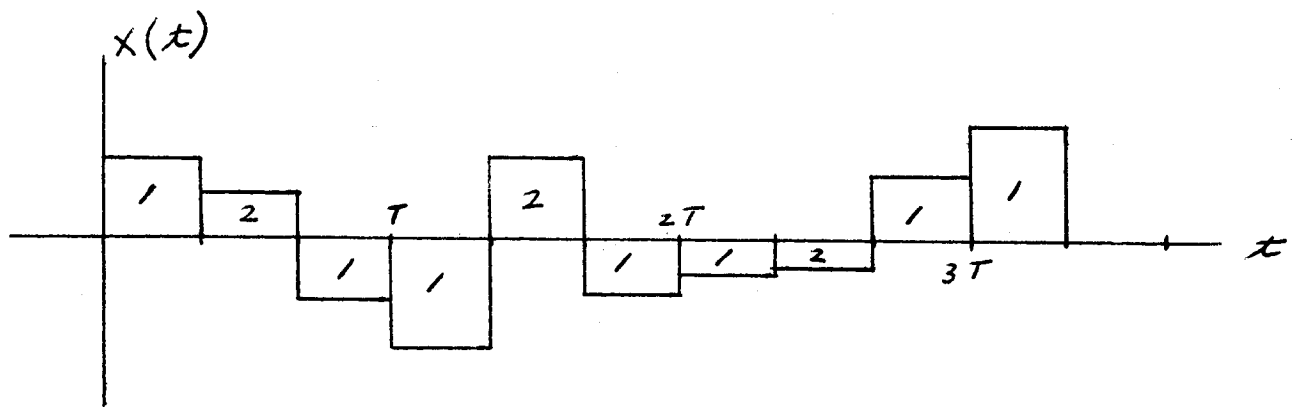


Fig. 15.3 Time Multiplexing of signals with unequal bandwidths.

It might be supposed that, since each data signal is only being observed for an infinitesimal fraction of the time, the bandwidth requirements could be considerably reduced. Actually this is not the case, for recall that the bandwidth B necessary to transmit a signal which exhibits variations of interest on the order to every τ seconds must satisfy the inequality $B \geq 1/2\tau$. The sampling theorem states that a signal of bandwidth W must be sampled at least every $\tau = 1/2W$ seconds. If this amplitude were transmitted as a pulse, the pulse could last only T seconds before the next sample must be sent. Thus the pulse width cannot be greater than $\tau = T = 1/2W$ and, hence, the bandwidth occupancy must be, at least $B = \frac{1}{2\tau} = W$ the bandwidth of the signal. The same comment applies to multiplexed signals; frequency multiplexed signals require a bandwidth at least as great as the sum of the bandwidth of the individual signal

$$W = \sum_i W_i \quad (15.2)$$

(Actually, the bandwidth defined in Eq. (15.1) is about twice this value. However, it will be recalled that by employing single sideband transmission the bandwidth may be halved without any loss of information. Thus, in the case of single-sideband frequency multiplexing the above statement holds. This could be done, of course, only at the expense of additional equipment). A time multiplexed system involving N data sources has only T/N seconds per pulse as was seen in Fig. 15.2. Thus, the bandwidth is increased by a factor of N in the case of equal signal bandwidths $W_i = W$; ($\tau = \frac{1}{2NW}$, $B = NW$).

Similarly, the frequency multiplexed signal bandwidth has increased by the same factor under the same condition as seen from Eq. (15.2). If the bandwidths are not equal, the frequency multiplexed signals can be spaced more efficiently, in general, since there is no necessity that the sampling rates be integrally related. However this advantage is offset by the necessity of single-sideband multiplexing to avoid increasing the bandwidth by a factor of two.

P. Pulse Modulation Systems and Matched Filtering

In the sections that follow a number of pulse modulation methods will be discussed. In order to simplify the discussion, it will be assumed that the data input to the transmitter consists of a sequence of samples at some average rate, say R samples per second. Thus each sample has $T = 1/R$ seconds in which to be transmitted. It is unimportant whether this sequence comes from one source or is the time multiplexed output from a number of sources. A pulse modulation system involves the transmission of a particular waveform $m(x)$ representing the sample in question for a period of time T seconds. The transmitted signal therefore, is allowed to change form only every T seconds.

Before proceeding to discuss various pulse modulating schemes in more detail, it is of interest to consider the generic form of the demodulators for pulse modulation. The pulse modulated signal, as observed, is characterized by a waveform $m_i(x)$ which is transmitted without interruption for the time interval $VT < x < (V+1)T$. There may be a finite or a continuous number of waveforms $m_i(x)$ which can be transmitted during each interval. (If the signal can assume a continuum of amplitudes for example, i may be infinite; if it is only desired to know the value of the amplitude to say, three decimal places, or if there are only a finite number of amplitudes possible, i may be finite). The set of waveforms $\{m_i(x)\}$ which can be transmitted is known at the receiver. The received signal will be perturbed by noise so that it will generally not match exactly any of the signals in this set. Suppose the signal $m_j(x)$ is transmitted and the signal $y(x) = A m_j(x) + n(x)$ is received where A is assumed to be a known constant. Hopefully $y(x)$ will resemble $A m_j(x)$ more closely than it resembles any of the other signals in the set $\{A m_i(x)\}$. If not, there is little chance of making a correct decision as to the transmitted signal. An ideal receiver would then have knowledge of the entire set $\{A m_i(x)\}$ and determine that waveform which most nearly resembles the received signal.

But how do we measure how closely one waveform resembles another? One method might be to determine the average value of the difference between the received waveform and each of the possible waveforms

$$\epsilon_j \triangleq \frac{1}{T} \int_{\nu T}^{(\nu+1)T} [y(x) - A m_j(x)] dx$$

and estimate the received signal as that corresponding to the ϵ_j with minimum absolute value. This would not be particularly effective, of course, since $y(x)$ could be much greater than $A m_j(x)$ over part of the interval, and must less over the remainder of the interval, and still yield a value of $\epsilon_j \approx 0$. An approach which avoids this difficulty (and in fact, one which can be shown to be optimum for many situations, including that normally encountered in space communications, is to determine the mean squared error;

$$\epsilon_j^2 \triangleq \frac{1}{T} \int_{\nu T}^{(\nu+1)T} [y(x) - A m_j(x)]^2 dx$$

If $\min_k \epsilon_k^2 = \epsilon_j^2$ the receiver concludes that the received signal was $m_j(x)$. Note that

$$\epsilon_j^2 = \frac{1}{T} \int_{\nu T}^{(\nu+1)T} y^2(x) dx - \frac{2A}{T} \int_{\nu T}^{(\nu+1)T} y(x) m_j(x) dx + \frac{A^2}{T} \int_{\nu T}^{(\nu+1)T} m_j^2(x) dx$$

and since the first term on the right is independent of j and the last is just the power in the j^{th} signal, the optimum signal is that for which the quantity

$$\frac{1}{T} \int_{\nu T}^{(\nu+1)T} m_j(x) y(x) dx = \frac{A E_j}{2T}$$

is a maximum. Here E_j represents the energy in the j^{th} signal $m_j(x)$. The term

$$\int_{\nu T}^{(\nu+1)T} y(x) m_j(x) dx$$

is called the cross-correlation between the signals $y(t)$ and $m_i(t)$. Consequently such a receiver is frequently called a correlation receiver. Another common designation is a matched-filter receiver. The process, of course, is repeated for all integer values of ν , so long as signals are being received.

Q. Pulse Amplitude Modulation PAM

Perhaps the most obvious method for transmitting sampled data is pulse amplitude modulation. If the data sample is X_ν , the signal $A x_\nu \cos \omega_c t$ is transmitted, $\nu T < t < (\nu+1)T$, the received signal then becoming $y(t) = B x_\nu \cos \omega_c t$. As discussed in the previous section, the optimum detector forms the quantity

$$g_i \left[(\nu+1)T \right] - \frac{B}{2T} E_i$$

$$= \frac{1}{T} \int_{\nu T}^{(\nu+1)T} y(t) x_i \cos \omega_c t dt - \frac{B}{2T} \int_{\nu T}^{(\nu+1)T} 2 x_i^2 \cos^2 \omega_c t dt$$

for every possible amplitude x_i of the received signal, and selects the largest of those as the best estimate of the received signal. But a condition that the quantity $g_i \left[(\nu+1)T \right] - \frac{B}{2T} E_i$ be a maximum is that

$$\frac{d}{dx_i} \left\{ g_i \left[(\nu+1)T \right] - \frac{B}{2T} E_i \right\} = 0$$

*There will be in general, a phase and probably even a frequency shift between the transmitter and the receiver. However, this will cause no difficulty so long as the received phase and frequency are determined at the receiver and used to generate the local signals $m_i(t)$. This knowledge will be assumed here so that the phase and frequency shift can safely be ignored.

or that

$$\frac{1}{T} \int_{\nu T}^{(\nu+1)T} y(x) \cos \omega_c x \, dx = \frac{x_i B}{T} \int_{\nu T}^{(\nu+1)T} 2 \cos^2 \omega_c x \, dx = x_i B$$

where the carrier frequency f_c has been chosen to be some multiple of half the reciprocal of the pulse period T , $f_c = \frac{\omega_c}{2\pi} = \frac{k}{2T}$ for some integer k . The optimum estimate \hat{x}_ν of the amplitude of the received signal, then is

$$\hat{x}_\nu = \frac{1}{T} \int_{\nu T}^{(\nu+1)T} y(x) \cos \omega_c x \, dx$$

and the receiver is simply that illustrated schematically in Fig. 17.1.

It should be noted that this estimate requires knowledge of the amplitude B of the arriving signal, a parameter seldom known to the receiver. It is a parameter which can readily be estimated however.

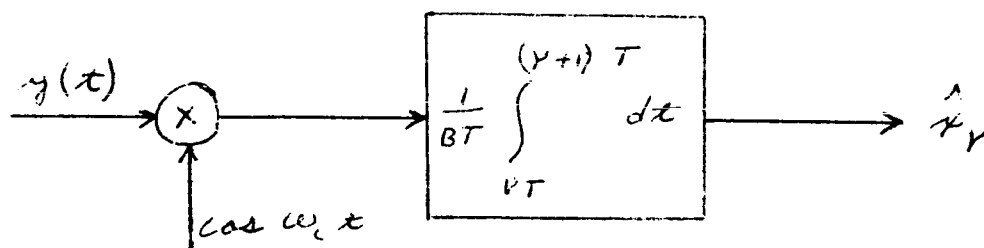


Fig. 17.1 A PAM detector.

To determine the signal-to-noise ratio at the output of a PAM detector, we note that the output signal may be written

$$\hat{x}_\nu = \frac{2}{T} \int_{\nu T}^{(\nu+1)T} x_\nu \cos^2 \omega_c x \, dx + \frac{1}{BT} \int_{\nu T}^{(\nu+1)T} n(x) \cos \omega_c x \, dx = x_\nu + m_\nu$$

where x_v is the output due to the received signal, and n_v that due to the noise. The output signal power therefore is just the power P_x in the modulating signal $x(t)$, (the fact that x_v is a time-quantized version of $x(t)$ does not alter its average power). As we have noted several times, multiplying white noise by a sinusoid does not alter its whiteness but alters its spectral density by a factor of $\frac{1}{2}$. It can be shown that the integrator acts as a filter with a noise bandwidth $B_n = \frac{1}{2T}$ so that the output noise power is $N_v / 2T B^2$ (the B^2 term due to the fact that the noise has been divided by B , the noise power therefore is reduced by the factor B^2). The output signal-to-noise ratio is consequently

$$\left(\frac{S}{N}\right)_{PAM} = \frac{B^2 P_x}{N_v \left(\frac{1}{2T}\right)} = \frac{B^2 P_x}{N_v B_n} \quad (15.3)$$

An interesting measure of the effective bandwidth of a signal is afforded by asking the question: How far in frequency must different communication channels be separated if the cross-modulation between any two of them is to be kept to an insignificant level? Suppose, in fact that a number of PAM channels were to be operated simultaneously at the carrier frequencies $\omega_i, i = 1, 2, \dots$. Then the effect of the signal in the i^{th} channel on the output of the j^{th} channel demodulator is simply

$$\begin{aligned} \frac{1}{BT} \int_{rT}^{(r+1)T} y_i(x) \cos \omega_j(x) &= \frac{1}{BT} \int_{rT}^{(r+1)T} x_r(x') \cos \omega_i x \cos \omega_j x \\ &= \frac{x_r(i)}{2BT} \left[\int_{rT}^{(r+1)T} \cos(\omega_i - \omega_j)x dx + \int_{rT}^{(r+1)T} \cos(\omega_i + \omega_j)x dx \right] \end{aligned}$$

which is identically zero if ω_i , ω_j , and $\omega_i - \omega_j$ are all non-zero integer multiples of the term π/T . Thus in order to keep the cross-modulation zero it is necessary to separate the channels in frequency by an amount $f_i - f_j = k/2T$, for any value of $k = 1, 2, \dots$. The effective bandwidth of each channel is therefore $B_{eff} = \frac{1}{2T} = W_{eff}$, where W is the bandwidth of the sampled signal. Consequently, Eq. (15.3) may be rewritten in terms of the effective bandwidth to yield

$$\left(\frac{S}{N}\right)_{PAM} = \frac{B^2 P_X}{N_0 B_{eff}}$$

Note that this is exactly the same relationship that was obtained for SSB modulation.

R. Phase-Shift Keyed Modulation (PSK)

Another rather common pulse modulation technique, called phase-shift keying, is to transmit the signal

$$\cos(\omega_c t + \pi_v), \quad vT < t < (v+1)T$$

to convey the data x_v . Here of course, $0 < \pi_v < 2\pi$ in order that there be no ambiguity at the receiver. Thus the phase, rather than the amplitude, conveys the information in a PSK system. The advantage of this method over PAM modulation rests in the fact that the amplitude of the signal remains constant. This is not an insignificant advantage in space telemetry since transmitters which work at a constant amplitude are considerably more efficient than those which must produce variable amplitudes.

The optimal PSK receiver, since the received signal energy is now independent of x_v , must form the integrals

$$\int_{vT}^{(v+1)T} y(t) \cos(\omega_c t + \pi_i) dt$$

for all $0 < \pi_i < 2\pi$ and select the largest. But this expression may be written

$$\begin{aligned} & \cos \pi_i \int_{vT}^{(v+1)T} y(t) \cos \omega_c t dt - \sin \pi_i \int_{vT}^{(v+1)T} y(t) \sin \omega_c t dt \\ &= X \cos \pi_i - Y \sin \pi_i. \end{aligned}$$

where $X = \int_{rT}^{(r+1)T} y(t) \cos \omega_c t \, dt$, and $Y = \int_{rT}^{(r+1)T} y(t) \sin \omega_c t \, dt$

The maximum of these with respect to x_i must satisfy the condition that

$$\frac{d}{dx_i} [X \cos x_i - Y \sin x_i] = 0$$

or that

$$x_i = \tan^{-1} \left(\frac{Y}{X} \right)$$

Thus the optimum estimate of x_r is $\hat{x}_r = \tan^{-1} \frac{Y}{X}$ and the optimum receiver is that depicted in Fig. 18.1

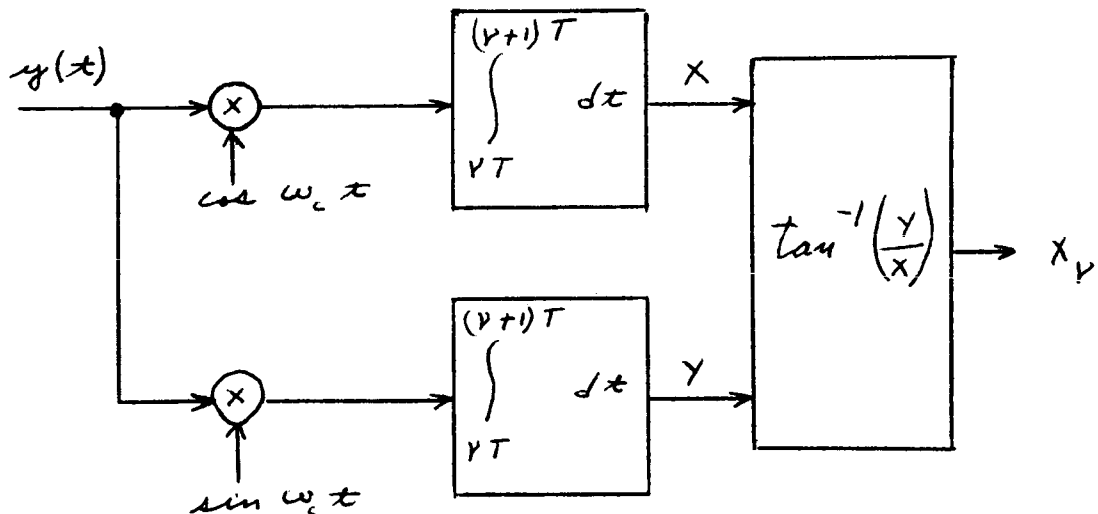


Fig. 18.1 A PSK detector.

It is observed that in the absence of noise,

$$y(t) = A \cos(\omega_c t + x_r)$$

and

$$\begin{aligned} X &= \int_{rT}^{(r+1)T} A \cos \omega_c t \cos(\omega_c t + x_r) dt \\ &= \frac{A}{2} \int_{rT}^{(r+1)T} [\cos x_r + \cos(2\omega_c t + x_r)] dt \\ &= \frac{AT}{2} \cos x_r \end{aligned}$$

while

$$\begin{aligned} Y &= \int_{rT}^{(r+1)T} A \sin \omega_c t \cos(\omega_c t + x_r) dt \\ &= \frac{A}{2} \int_{rT}^{(r+1)T} [\sin x_r + \sin(2\omega_c t + x_r)] dt \\ &= \frac{AT}{2} \sin x_r \end{aligned}$$

where it is again assumed that $\omega_c = \frac{\pi k}{T}$ for some integer k .

Hence

$$\hat{X}_r = \tan^{-1} \left[\frac{AT \sin x_r}{AT \cos x_r} \right] = x_r$$

and the estimate of the signal is exact in the absence of noise. Additionally, note that this system does not require knowledge of the signal amplitude for its operation. The evaluation of the performance will certainly depend on the amplitude of the received waveform. The analysis of the output signal-to-noise ratio for PSK is somewhat more involved than that for PAM and will not be carried out here. The results of such an analysis, however, would indicate performance approximately equal to that of PAM. In addition, it is easily verified that essentially the same comments concerning the spectrum of a PAM signal as well as its effective bandwidth occupancy apply equally to a PSK modulated signal. (The effective bandwidth occupancy of a PSK signal is actually twice that of a PAM signal).

S. Pulse Code Modulation (PCM)

In all of the modulation schemes discussed up to now, it has been possible to transmit any of a continuum of amplitudes (generally assumed to be bounded by finite values). But several observations strongly suggest that this is not necessary. In the first place, regardless of the use to which the receiver of the information intends to apply it, he can never use more than a finite number of significant digits for each sample. If for no other reason, the accuracy of the measuring device is always limited to some degree. Moreover, the noise encountered at the receiver insures that the data will contain some inaccuracies nullifying the meaningfulness of all but the most significant digits in the received samples. Thus, so far as the receiver is concerned, it is of little consequence whether or not the data is quantized in amplitude as well as in time. Because many of the measurements are inherently digital in nature anyway (the outputs of counters, for example) it may be of some advantage to quantize all the information so that it may all be processed uniformly on the spacecraft. But is there any advantage so far as the communication system itself is concerned, to quantizing all analogue signals before transmission? The answer is emphatically yes, as the following analysis of a pulse-code modulation system will illustrate.

Since all the data is to be quantized, the communication system need be concerned with only a finite number of signals $m_i(t)$. In particular, suppose the quantized information is represented in binary form. If the output of a device were quantized to four levels, for example, the possible outputs could

be represented by the four two-digit numbers

00
01
10
11

In general, if the output is quantized to M levels, each possible output can be represented by $\log_2 M$ (or the smallest integer greater than $\log_2 M$ if it is not an integer) binary symbols or bits. (Of course, bases other than the binary base could be used, but binary systems possess certain distinct advantages and are by far the most commonly encountered).

A pulse-code modulation (PCM) system is one in which each of the binary digits used to represent a data sample is individually transmitted as one of two possible waveforms. That is, if the bit is a one, some signal $m_1(t)$ is transmitted for a duration of T_B seconds; if it is a zero some other signal is transmitted for the same duration. Then the next bit is similarly transmitted, etc. Typically, $m_2(t) = -m_1(t)$. The PCM correlation receiver is shown in Fig. 19.1. Note that if $m_2(t)$ is equal

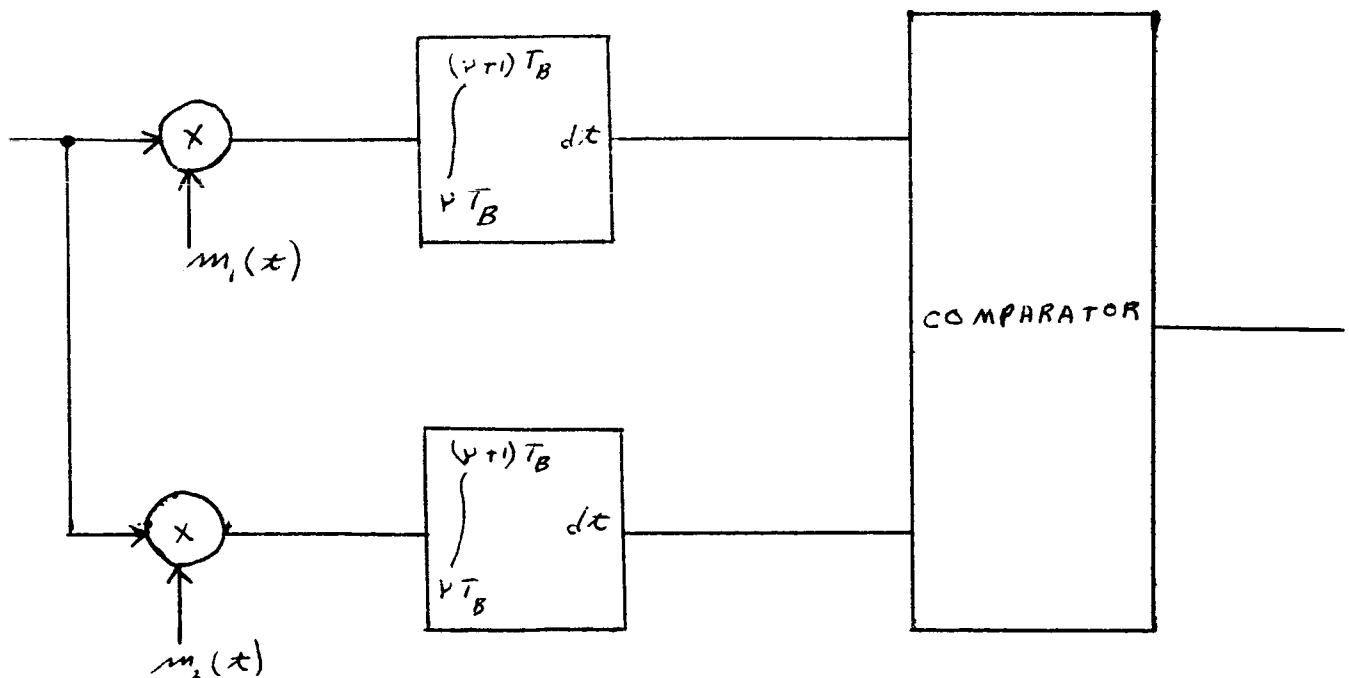


Fig. 19.1 An optimum PCM detector.

to $m_1(x)$ only one integrator is needed. The comparator then decides that a one was received if the integrator output is positive, and that a zero was received if it is negative.

To analyze the output signal-to-noise ratio using this modulation method, let us assume that the ratio of the received signal power to the noise power is sufficiently large so that it is most unlikely that $m_1(x)$ can be mistaken for $m_2(x)$ at the receiver, or conversely. The significance of this assumption will be investigated shortly. We observe that in this case the noise is just that due to the fact that the signal is quantized. That is, if the sampled signal is $s(x) = s(\nu T)$, $\nu T < x < (\nu+1)T$ the quantized signal $s_q(\nu T)$ is transmitted. Since no mistake is made at the receiver in identifying each of the digits representing $s_q(x)$ this signal is recovered exactly. Thus the mean-squared noise at the receiver is just

$$\frac{1}{T} \int_{\nu T}^{(\nu+1)T} [s(x) - s_q(x)]^2 dx = [s(\nu T) - s_q(\nu T)]^2$$

Suppose the signal is equally likely to assume any amplitude from $-a$ to a , and that the quantization is such that

$$s_q(\nu T) = \frac{2a(i + \frac{1}{2})}{M}$$

when

$$\frac{2ai}{M} < s(\nu T) < \frac{2a(i+1)}{M}, \quad i = -\frac{M}{2}, -\frac{M}{2} + 1, \dots, \frac{M}{2} - 1.$$

Then, while the signal varies between the extremes of $-a$ and a is equally likely to have any value in that range at any given time, the quantization error varies between $\frac{-a}{M}$ and $\frac{a}{M}$ and is equally likely to assume any value in that range. Recalling the interpretation given to the signal-to-noise ratio, we see that we have reduced the uncertainty in the amplitude of the signal by a factor of M upon the receipt of the quantized signal $s_q(x)$. If signals are to be transmitted every T_0 seconds, and each one is to be quantized into $M=2^m$ levels

then each bit has only

$$\frac{T}{B} = \frac{T_c}{\log_2 M} = \frac{T_c}{m}$$

seconds in which to be transmitted. Recall that the previous pulse modulation systems have shown a strong dependence on the amount of time spent in transmitting any particular waveform. This dependence is equally true here. If, for example, $m_2(t) = -m_1(t)$ and the average power in the signal $m_1(t)$ at the receiver is \bar{B}^2 and the noise spectral density is N_0 , then it is usually sufficient to require that

$$\frac{\bar{B}^2 T_B}{2 N_0} \geq 5 \quad (19.1)$$

in order that the two signals $m_1(t)$ and $-m_1(t)$ can be distinguished reliably enough at the receiver to justify the assumption made in this derivation.

Again, as in the case of PAM, the effective bandwidth necessary to transmit a pulse of time duration T_B is

$$B_{eff} = \frac{1}{2 T_B}$$

Thus condition 19.1 becomes

$$\frac{\bar{B}^2}{N_0 B_{eff}} \geq 20$$

and is consequently analogous to the condition in section 18 for the FM threshold.